ARCHIVE
for
RATIONAL MECHANICS
and
ANALYSIS

UNIVERSITY OF HAWAII
LIBRARY
AUG 24 '61

Edited by

C. TRUESDELL

Volume 8, Number 1

SPRINGER-VERLAG
BERLIN · GÖTTINGEN · HEIDELBERG
(Postverlagsort Berlin · 4. 7. 1961)

801
7

Mechanicam vero duplicem Veteres constituerunt: Rationalem *quae per Demonstrationes accurate procedit, & Practicam. Ad practicam spectant Artes omnes Manuales, a quibus utique* Mechanica *nomen mutuata est. Cum autem Artifices parum accurate operari soleant, fit ut* Mechanica *omnis a* Geometria *ita distinguatur, ut quicquid accuratum sit ad* Geometriam *referatur, quicquid minus accuratum ad* Mechanicam. *Attamen errores non sunt Artis sed Artificum. Qui minus accurate operatur, imperfectior est Mechanicus, & si quis accuratissime operari posset, hic foret Mechanicus omnium perfectissimus.* NEWTON

La généralité que j'embrasse, au lieu d'éblouïr nos lumieres, nous découvrira plutôt les véritables loix de la Nature dans tout leur éclat, & on y trouvera des raisons encore plus fortes, d'en admirer la beauté & la simplicité. EULER

. . . *ut proinde his paucis consideratis tota haec materia redacta sit ad puram Geometriam, quod in physicis & mechanicis unice desideratum.* LEIBNIZ

The ARCHIVE FOR RATIONAL MECHANICS AND ANALYSIS nourishes the discipline of mechanics as a deductive, mathematical science in the classical tradition and promotes pure analysis, particularly in contexts of application. Its purpose is to give rapid and full publication to researches of exceptional moment, depth, and permanence.

Each memoir must meet a standard of rigor set by the best work in its field. Contributions must consist largely in original research; on occasion, an expository paper may be invited.

English, French, German, Italian, and Latin are the languages of the Archive. Authors are urged to write clearly and well, avoiding an excessively condensed or crabbed style.

Manuscripts intended for the Archive should be submitted to an appropriate member of the Editorial Board.

The ARCHIVE FOR RATIONAL MECHANICS AND ANALYSIS appears in numbers struck off as the material reaches the press; five numbers constitute a volume. Subscriptions may be entered through any agent. The price is DM 96.— per volume.

Notice is hereby given that for all articles published exclusive rights in all languages and countries rest with Springer-Verlag. Without express permission of Springer-Verlag, no reproduction of any kind is allowed.

For each paper 75 offprints are provided free of charge.

5-856-24A

Poiseuille Flow of Certain Anisotropic Fluids

J. L. Ericksen

1. Introduction

Simple shearing flow of relatively simple, incompressible anisotropic fluids is analyzed by Ericksen [*1*]. For a certain class of these, the predicted behavior is like that of the so-called Bingham materials. That is, the steady state shear stress is a linear but not homogeneous function of the rate of shear. The purpose of this paper is to present analyses indicating how these might behave in Poiseuille flow, *i.e.* in steady flow through an infinitely long circular pipe. It turns out that there are infinitely many solutions which seem more or less acceptable. Acceptance of a plausibility argument rules out all but one on the grounds that the remainder represent unstable situations. That which remains yields results entirely similar to those deduced by Buckingham [*2*] and Reiner [*3*] for Poiseuille flow of Bingham materials. Though this might be expected, it is not entirely obvious *a priori*. In the one dimensional descriptions of Bingham materials used by Buckingham and Reiner, or in the three dimensional formulation introduced by Oldroyd [*4*], two sets of constitutive equations are used. One set, applying at low stresses, describes an ordinary elastic material. The other, applying at high stresses, involves a viscosity similar to that which characterizes Newtonian fluids. A scalar yield criterion is used as a criterion for determining which set applies. The theory used here employs a single set of constitutive equations and no yield criterion. One consequence is that the reasoning used in analyzing Poiseuille flow of Bingham materials is not directly applicable to the materials here considered.

2. Equations of Motion

It is a straightforward matter to refer the equations proposed by Ericksen [*1*, eqs. 19 and 20] to cylindrical coordinates (r, ϑ, z), then to introduce the assumptions that the velocity field is of the form

$$\frac{dr}{dt} = \frac{d\vartheta}{dt} = 0, \qquad \frac{dz}{dt} = f(r), \tag{2.1}$$

and that all quantities are independent of ϑ. One obtains the stress-deformation relations

$$\begin{aligned} t_{rr} &= -p + X n_r^2 + 2\mu_3 f' n_r n_z, \\ t_{\vartheta\vartheta} &= -p + X n_\vartheta^2, \\ t_{zz} &= -p + X n_z^2 + 2\mu_3 f' n_r n_z, \\ t_{rz} &= X n_r n_z + [\mu_3(1 - n_\vartheta^2) + \mu] f', \\ t_{r\vartheta} &= X n_r n_\vartheta + \mu_3 f' n_\vartheta n_z, \\ t_{\vartheta z} &= X n_\vartheta n_z + \mu_3 f' n_r n_\vartheta, \\ X &\equiv \mu_1 + \mu_2 f' n_r n_z. \end{aligned} \tag{2.2}$$

Here t_{rr}, $t_{\vartheta\vartheta}$, *etc.*, are physical components of the symmetric stress tensor, p is an arbitrary pressure and (n_r, n_ϑ, n_z) are physical components of a unit vector pointing out a preferred direction in the fluid. Greek letters denote material constants. The vector satisfies the differential equations

$$\begin{aligned} 2\frac{dn_r}{dt} &= f'\, n_z(\lambda - 1 - 2\lambda n_r^2), \\ \frac{dn_\vartheta}{dt} &= -\lambda f'\, n_r n_\vartheta n_z, \\ 2\frac{dn_z}{dt} &= f'\, n_r(\lambda + 1 - 2\lambda n_z^2), \end{aligned} \tag{2.3}$$

where

$$\frac{d}{dt} = \frac{\partial}{\partial t} + f(r)\frac{\partial}{\partial z}$$

is the material derivative operator. The fluid behaves as a Bingham material provided

$$|\lambda| > 1, \qquad \mu_1 \neq 0, \tag{2.4}$$

which we assume.

Flows of the form (2.1) are accelerationless. In the absence of body forces, the stresses must thus satisfy the equations of static equilibrium, which reduce to

$$\begin{aligned} &\frac{\partial t_{rr}}{\partial r} + \frac{\partial t_{rz}}{\partial z} + \frac{1}{r}(t_{rr} - t_{\vartheta\vartheta}) = 0, \\ &\frac{\partial t_{r\vartheta}}{\partial r} + \frac{\partial t_{z\vartheta}}{\partial z} + \frac{2}{r} t_{r\vartheta} = 0, \\ &\frac{\partial t_{rz}}{\partial r} + \frac{\partial t_{zz}}{\partial z} + \frac{1}{r} t_{rz} = 0. \end{aligned} \tag{2.5}$$

From (2.1), $df'/dt = 0$, so f' may be treated as a constant in integrating (2.3). Equations of precisely the same form are analyzed by Ericksen [*1*, *5*], from which we may conclude that, when (2.4) holds, essentially all solutions of (2.3) approach the solution

$$n_r = \sin\varphi, \qquad n_\vartheta = 0, \qquad n_z = \cos\varphi \tag{2.6}$$

as $t \to \infty$. Here φ is a constant angle determined by the conditions that

$$\tan^2\varphi = (\lambda - 1)/(\lambda + 1), \tag{2.7}$$

and that it lie in the range

$$0 < \varphi < \pi/4 \quad \text{when} \quad f' > 0, \quad \lambda > 1, \tag{2.8}$$

$$3\pi/4 < \varphi < \pi \quad \text{when} \quad f' < 0, \quad \lambda > 1, \tag{2.9}$$

$$\pi/4 < \varphi < \pi/2 \quad \text{when} \quad f' < 0, \quad \lambda < -1, \tag{2.10}$$

$$\pi/2 < \varphi < 3\pi/4 \quad \text{when} \quad f' > 0, \quad \lambda < -1. \tag{2.11}$$

The theory treats $\boldsymbol{n}$ and $-\boldsymbol{n}$ as indistinguishable, which means that there is no loss in generality in taking $0 \leqq \varphi \leqq \pi$, as is done here.

3. Steady State Flow

We use the stable, steady state solution of (2.3) given by (2.6)—(2.11), proceeding under the assumption that $f' \neq 0$. Using (2.2), we then find that

$$t_{rr} = -p + \mu_1 \sin^2\varphi + A f', \tag{3.1}$$

$$t_{\vartheta\vartheta} = -p, \tag{3.2}$$

$$t_{zz} = -p + \mu_1 \cos^2\varphi + B f', \tag{3.3}$$

$$t_{rz} = C + D f', \tag{3.4}$$

$$t_{r\vartheta} = t_{\vartheta z} = 0, \tag{3.5}$$

where A, B, C and D are constants given by

$$2A = (\mu_2 \sin^2\varphi + 2\mu_3) \sin 2\varphi, \tag{3.6}$$

$$2B = (\mu_2 \cos^2\varphi + 2\mu_3) \sin 2\varphi, \tag{3.7}$$

$$2C = \mu_1 \sin 2\varphi, \tag{3.8}$$

$$4D = 4(\mu_3 + \mu) + \mu_2 \sin^2 2\varphi. \tag{3.9}$$

Since $t_{rz} \to C$ as $f' \to 0$, C is interpretable as an apparent yield shear stress. Similarly, D is an apparent fluidity. It should be noted that the constants A, B and C reverse sign when f' does, so places where $f' = 0$ must be treated with care. For the moment, we exclude these. It is then a simple matter to integrate (2.5) to obtain

$$p = \mu_1 \sin^2\varphi + A f' + \int \frac{1}{r} (\mu_1 \sin^2\varphi + A f')\, dr - E z, \tag{3.10}$$

$$t_{rz} = C + D f' = -E r/2 + F/r, \tag{3.11}$$

where E and F are constants of integration.

Physically, we expect the stresses to do non-negative work in deforming the fluid,

$$t_{rz} f' = (C + D f') f' \geqq 0. \tag{3.12}$$

From (3.8), (3.9) and (2.7)—(2.11), we find that this requires that

$$\mu_1 \geqq 0, \qquad \mu_2(\lambda^2 - 1) + 4\lambda^2(\mu_3 + \mu) \geqq 0 \quad \text{when} \quad \lambda > 1, \tag{3.13}$$

$$\mu_1 \leqq 0, \qquad \mu_2(\lambda^2 - 1) + 4\lambda^2(\mu_3 + \mu) \geqq 0 \quad \text{when} \quad \lambda < -1. \tag{3.14}$$

We assume that the strict inequalities hold in (3.12) and (3.13).

It is easily seen that our assumptions rule out the possibility of having a solution devoid of singularities at $r = 0$. Previously mentioned analyses of Bingham materials suggest that, near the center of a pipe, plug flow occurs. That is, a central core moves as a rigid body. We now explore this possibility.

4. Plug Flow

We seek a steady solution of the basic equations with $f = f_0 = \text{const.}$ which is devoid of singularities at $r = 0$. From (2.3), the material derivative of $\boldsymbol{n}$ must vanish, which implies that $\boldsymbol{n}=\boldsymbol{n}(r, \vartheta, z-f_0 t)$. Dependence on ϑ was excluded earlier and, except when $f_0=0$, dependence on $z-f_0 t$ is ruled out by the assumption of steadiness. We thus assume that

$$n_r = n_r(r), \quad n_\vartheta = n_\vartheta(r), \quad n_z = n_z(r). \tag{4.1}$$

From (2.2) and (2.5), we then have

$$r^2 t_{r\vartheta} = \mu_1 n_r n_\vartheta = G = \text{const.}$$

For $t_{r\vartheta}$ to remain finite as $r \to 0$, $G=0$, so that

$$n_r n_\vartheta = 0. \tag{4.2}$$

From (2.2),

$$n_r = 0 \Rightarrow t_{rz} = 0. \tag{4.3}$$

As will be discussed later, (4.3) makes the possibility $n_r=0$ uninteresting for pipe flow. Since $\boldsymbol{n}$ is a unit vector, (4.1) and (4.2) leave us with the possibly interesting alternative

$$n_r = \sin\psi, \quad n_\vartheta = 0, \quad n_z = \cos\psi, \tag{4.4}$$

where ψ is any function of r. From (2.2), we then have

$$t_{rr} = -p + \mu_1 \sin^2\psi, \tag{4.5}$$

$$t_{\vartheta\vartheta} = -p, \tag{4.6}$$

$$t_{zz} = -p + \mu_1 \cos^2\psi, \tag{4.7}$$

$$t_{rz} = \tfrac{1}{2}\mu_1 \sin 2\psi, \tag{4.8}$$

$$t_{r\vartheta} = t_{\vartheta z} = 0. \tag{4.9}$$

Integrating (2.5) subject to the requirement that t_{rz} remain finite as $r \to 0$, we obtain

$$p = \mu_1 \sin^2\psi + \mu_1 \int \frac{1}{r} \sin^2\psi \, dr - H z, \tag{4.10}$$

$$2 t_{rz} = \mu_1 \sin^2\psi = -H r, \tag{4.11}$$

where H is a constant of integration. From (4.9), we should have $\psi \to 0$ as $r \to 0$ to keep the normal stresses finite as $r \to 0$. Because of the indistinguishability of $\boldsymbol{n}$ and $-\boldsymbol{n}$, there is no loss in generality in taking $0 \leqq \psi < \pi$. Solving (4.11) for $\sin\psi$ subject to these conditions, we obtain

$$\sin^2\psi = \tfrac{1}{2}\left[1 - \sqrt{1 - (H r/\mu_1)^2}\right], \tag{4.12}$$

with

$$0 \leqq \psi \leqq \pi/4 \quad \text{if} \quad H/\mu_1 < 0, \tag{4.13}$$

$$3\pi/4 \leqq \psi \leqq \pi \quad \text{if} \quad H/\mu_1 > 0. \tag{4.14}$$

This gives solutions of the desired type, valid for

$$0 \leqq r \leqq r_1 = |\mu_1/H|. \tag{4.15}$$

5. Combined Flow

We consider the possibility of using a solution of the type discussed in §3 for $r \geqq r_2$, a solution of the type discussed in §4 for $r \leqq r_2$, requiring that the velocity vectors and stress vectors should match at $r = r_2$. That is, f, t_{rr}, $t_{r\vartheta}$ and t_{rz} should be continuous at $r = r_2$. For the moment, r_2 represents any value of r in the range $0 \leqq r \leqq r_1$, where r_1 satisfies (4.15). From (3.5) and (4.9), $t_{r\vartheta}$ is automatically continuous, being zero according to both solutions.

For plug flow, we have from (4.5), (4.10) and (4.12)

$$t_{rr}(r) = -\frac{1}{2}\int\limits_0^r \frac{1}{x}\left[1 - \sqrt{1 - (H\,x/\mu_1)^2}\right] d\,x + H\,z + K, \qquad r \leqq r_2 \tag{5.1}$$

where K is a constant of integration. From (3.1) and (3.10), we may write

$$t_{rr}(r) = -\int\limits_{r_2}^r \frac{1}{x}\,(\mu_1 \sin^2\varphi + A\,f')\,d\,x + E\,z + L, \qquad r \geqq r_2. \tag{5.2}$$

For (5.1) and (5.2) to agree for $r = r_2$, z arbitrary, we must have

$$E = H, \tag{5.3}$$

$$L = -\frac{1}{2}\mu_1 \int\limits_0^{r_2} \frac{1}{x}\left[1 - \sqrt{1 - (H\,x/\mu_1)^2}\right] d\,x + K. \tag{5.4}$$

Using (5.3), (3.11) and (4.11), we see that t_{rz} is continuous at $r = r_2$ if and only if

$$F = 0 \tag{5.5}$$

which means that

$$t_{rz} = -H\,r/2 \tag{5.6}$$

throughout. From (3.11) and (5.6),

$$f = -\frac{1}{4D}\,(H\,r^2 + 4\,C\,r) + M, \tag{5.7}$$

where M is an arbitrary constant. For f to be continuous throughout,

$$f_0 = -\frac{1}{4D}\,(H\,r_2^2 + 4\,C\,r_2) + M. \tag{5.8}$$

As long as we satisfy these conditions, we obtain a combination of solutions which is physically acceptable for all values of r. This would leave us with an overabundance of solutions to the Poiseuille flow problem. We introduce a plausibility argument which rules out most of the possibilities on the grounds that they describe unstable situations.

From the plug flow solution, we can calculate $\bar{\boldsymbol{n}}$, the limiting value of $\boldsymbol{n}$ for $r \to r_2$, $r < r_2$, as a function of r_2,

$$\bar{n}_r = \sin\psi(r_2), \qquad \bar{n}_\vartheta = 0, \qquad \bar{n}_z = \cos\psi(r_2). \tag{5.9}$$

For the flow external to r_2, we have from (5.7) the limiting velocity gradient

$$f'(r_2) = -\frac{1}{2D}\,(H\,r_2 + 2C), \tag{5.10}$$

which will in general tend to rearrange the structure represented by $\boldsymbol{n}$. That is, if we regard $\bar{\boldsymbol{n}}$ as an initial value of $\boldsymbol{n}$ in (2.3), we see that $f'(r_2)$ would cause $\boldsymbol{n}$ to change in accordance with

$$2\frac{dn_r}{dt}=f'(r_2)\,n_z(\lambda-1-2\lambda n_r^2),$$
$$2\frac{dn_z}{dt}=f'(r_2)\,n_r(\lambda+1-2\lambda n_z^2),$$

n_ϑ remaining zero. We expect this to make the interface unstable in general. Exceptions might occur if

$$\psi(r_2)=\varphi, \tag{5.11}$$

when $\boldsymbol{n}$ coincides with the stable, steady state value of $\boldsymbol{n}$ in the flow exterior to $r=r_2$, or if

$$f'(r_2)=0 \tag{5.12}$$

which implies that $d\boldsymbol{n}/dt=0$. If (5.11) holds, (4.11) and (3.8) yield

$$r_2=-(\mu_1/H)\sin 2\psi(r_2)=-(\mu_1/H)\sin 2\varphi=-2C/H, \tag{5.13}$$

whence it follows from (5.10) that (5.12) holds. Hence (5.12) must hold and, at the interface

$$t_{rz}=-\tfrac{1}{2}H\,r_2=C.$$

Thus the shear stress takes on its apparent yield value at the interface. This is in accord with analyses of Poiseuille flow in Bingham materials, wherein it is assumed or deduced from other assumptions that, at the boundary of the rigidly moving plug, the velocity gradient vanishes, the shear stress taking on its yield value.

If (5.12) holds, (5.13) must hold, from which we may infer that either (5.11) holds or

$$\psi(r_2)=\pi/2-\varphi. \tag{5.14}$$

Noting (4.13) and (4.14), then referring to (2.8)—(2.11), we see that (5.11) but not (5.13) can hold if $\lambda>1$, whereas (5.13) but not (5.11) can hold if $\lambda<-1$.

The rather crude argument used here hardly settles the stability question for these flows. Among other things, it says nothing concerning the stability of the noted exceptional cases. It seems probable that, under some conditions, this exception describes a stable situation, but we leave this question to the future.

6. Poiseuille Flow

We consider using the solutions developed to describe steady flow through an infinitely long, stationary circular pipe of radius R, to which the fluid adheres, regarding as given the pressure head

$$H=-\partial p/\partial z=\partial t_{zz}/\partial z.$$

Assume the consequences of the plausibility argument used in §5 and suppose H is sufficiently large to make

$$r_2=-2C/H<R. \tag{6.1}$$

Using the adherence condition to evaluate the constant M in (5.7) and (5.8), we have

$$f = -\frac{1}{4D}(r-R)\left[H(r+R)+4C\right], \qquad r \geqq r_2,$$

$$f = f_0 = -\frac{1}{4D}(r_2-R)\left[H(r_2+R)+4C\right], \qquad r \leqq r_2,$$

where C and D are given by (3.8) and (3.9), *viz*

$$2C = \mu_1 \sin 2\varphi, \qquad 4D = \mu_2 \sin^2 2\varphi + 4(\mu_3+\mu).$$

To determine the appropriate value of φ, we note that (5.13) requires that

$$-\mu_1 H \sin 2\varphi > 0,$$

Then use (3.13) or (3.14) and (2.7)—(2.11). For a given material, we thus uniquely determine the velocity distribution for the combined flow as long as (6.1) holds. With the understanding that C and D are interpretable as yield stress and fluidity, it is indistinguishable from the velocity distribution for a Bingham material. We thus obtain the same expression for mass flux in terms of pressure head, given *e.g.* by REINER [*3*].

If the pressure head is sufficiently low, (6.11) must fail, the radius r_2 then lying outside the pipe. A reasonable interpretation is that the material remains at rest. In other words, the plug flow solution with $f_0=0$ applies. From (4.5) to (4.14) and from (5.1) we then obtain the stress

$$\begin{aligned} t_{rr} &= Y, \\ t_{\vartheta\vartheta} &= Y - \tfrac{1}{2}\mu_1\left[1-\sqrt{1-(Hr/\mu_1)^2}\right], \\ t_{zz} &= Y + \mu_1\sqrt{1-(Hr/\mu_1)^2}, \\ t_{rz} &= -\tfrac{1}{2}Hr, \\ t_{r\vartheta} &= t_{\vartheta z} = 0, \end{aligned} \tag{6.2}$$

where

$$Y = -\frac{1}{2}\mu_1 \int_0^r \frac{1}{x}\left[1-\sqrt{1-(Hx/\mu_1)^2}\right]dx + Hz + K.$$

when flow occurs, these stresses obtain for $0 \leqq r \leqq r_2$. For $r_2 \leqq r \leqq R$, we then similarly obtain

$$\begin{aligned} t_{rr} &= Z, \\ t_{\vartheta\vartheta} &= Z - \mu_1 \sin^2\varphi + \frac{A}{2D}(Hr+2C), \\ t_{zz} &= Z + \mu_1 \cos 2\varphi + \frac{A-B}{2D}(Hr+2C), \\ t_{rz} &= -\tfrac{1}{2}Hr, \\ t_{r\vartheta} &= t_{\vartheta z} = 0 \end{aligned} \tag{6.3}$$

where A, B, C, D are constants given by (3.6)—(3.9),

$$Z = -\int_{r_2}^{r} \frac{1}{x}(\mu_1 \sin^2\varphi + Af')\,dx + Ez + L, \tag{6.4}$$

and L is given by (5.4). Carrying out the integration in (6.4), we obtain

$$Z = (A\,C/D - \mu_1 \sin^2 \varphi) \ln (r/r_2) + \frac{A\,H}{2D} (r - r_2) + E\,z + L\,.$$

Except for K, which remains arbitrary, all constants are uniquely determined as functions of H. Changing K changes the stress by a uniform hydrostatic pressure, which has no effect on incompressible materials.

The stresses given by (6.2) or (6.3) differ from those predicted by Oldroyd's [4] equations for Bingham materials, which imply that $t_{rr} = t_{\vartheta\vartheta} = t_{zz}$. However, in terms of quantities generally discussed in connection with Poiseuille flow of Bingham materials, the two theories yield indistinguishable predictions.

For Poiseuille flow, the possibility indicated in (4.3), *i.e.* $n_r = 0$, $n_\vartheta \neq 0$, seems uninteresting for several reasons. One is that it follows from (3.13) and (3.14) that it is impossible to match the shear stress t_{rz}, which vanishes, with that given by (3.11) at some value of r. Another is that it gives rise to a couple, associated with the shear stress $t_{\vartheta z} = \mu_1 n_\vartheta n_z$, which is inappropriate for Poiseuille flow.

This work was supported by a grant from the Office of Ordnance Research, U. S. Army.

References

[1] Ericksen, J. L.: Transversely isotropic fluids. Koll. Z. **173**, 117–122 (1960).
[2] Buckingham, E.: On plastic flow through capillary tubes. Proc. Amer. Soc. Test. Mater. **21**, 1154–1156 (1921).
[3] Reiner, M.: Über die Strömung einer elastischen Flüssigkeit durch eine Kapillare. Koll. Z. **39**, 80–87 (1926).
[4] Oldroyd, J. G.: A rational formulation of the equations of plastic flow for a Bingham solid. Proc. Cambr. Phil. Soc. **43**, 100–105 (1947).
[5] Ericksen, J. L.: A vorticity effect in anisotropic fluids. J. Polymer Sci. **47**, 327–331 (1960).

Mechanics Department
The Johns Hopkins University
Baltimore 18, Maryland

(Received March 23, 1961)

The Finite Flexure of an Aeolotropic Elastic Cuboid

J. E. ADKINS & A. E. GREEN

1. Introduction

In a recent book by GREEN & ADKINS (1960) a part* of Chapter II is devoted to the flexure problem for rectilinear and curved cuboids of aeolotropic materials. Some of this work is a special case of a number of general problems considered by ADKINS (1955) and, unfortunately, in re-writing the work for the book the wrong impression** has been given that some of the results for the particular flexure problem which is considered apply to a body possessing a completely general strain energy function W. If τ^{ij} is the stress tensor referred to a curvilinear system of coordinates ϑ^i, measured per unit area of the deformed body, and we take ϑ^i to coincide with cylindrical polar coordinates (r, ϑ, y_3) in the deformed body, so that

$$(\vartheta^1, \vartheta^2, \vartheta^3) = (r, \vartheta, y_3), \tag{1.1}$$

then τ^{12}, τ^{13} must, in general, vanish if the equations of equilibrium and boundary conditions are to be satisfied. In the problems considered by GREEN & ADKINS this implies some restrictions on the form of W (see below) and although the restrictions were in fact used throughout their work, some statements to the contrary are made, especially in §2.8.

In a paper mainly concerned with general results in terms of a parametric representation, ADKINS (1955) examined the combined extension, shear and flexure of a cuboid and gave some results which, in principle, were valid for a general homogeneous aeolotropic body. We briefly re-examine this problem here in order to re-emphasize the conditions under which it may be regarded as solved for a material having a general strain energy function.

2. Notation and formulae

We quote the notation and formulae of GREEN & ADKINS (1960). The strain energy W of a homogeneous aeolotropic body can be expressed as a polynomial

$$W = W(e_{ij}) \tag{2.1}$$

in the components of strain e_{ij}, where

$$2e_{ij} = \frac{\partial y^r}{\partial x^i}\frac{\partial y^r}{\partial x^j} - \delta_{ij}. \tag{2.2}$$

---

* §§ 2.8, 2.9, 2.10, 2.14—2.18.

** We are indebted to Professor TRUESDELL for drawing our attention to this.

In (2.2) x^i, y^i are the coordinates of points in the undeformed and deformed bodies respectively, referred to the same rectangular Cartesian system of axes. The stress tensor τ^{ij}, defined in § 1, is given by

$$\tau^{ij} = \frac{1}{2\sqrt{I_3}}\left(\frac{\partial W}{\partial e_{rs}} + \frac{\partial W}{\partial e_{sr}}\right)\frac{\partial \vartheta^i}{\partial x^r}\,\frac{\partial \vartheta^j}{\partial x^s}, \tag{2.3}$$

for a compressible body, where

$$I_3 = |2e_{rs} + \delta_{rs}|. \tag{2.4}$$

If the body is incompressible, $I_3 = 1$ and

$$\tau^{ij} = \frac{1}{2}\left(\frac{\partial W}{\partial e_{rs}} + \frac{\partial W}{\partial e_{sr}}\right)\frac{\partial \vartheta^i}{\partial x^r}\,\frac{\partial \vartheta^j}{\partial x^s} + p\,G^{ij} \tag{2.5}$$

where p is a scalar function of ϑ^i and G^{ij} is the contravariant metric tensor of the curvilinear coordinates ϑ^i in the deformed body. The equations of equilibrium, in the absence of body forces, are

$$\tau^{ij}\|_j = 0, \tag{2.6}$$

the double line denoting differentiation with respect to curvilinear coordinates ϑ^i in the deformed body.

3. Flexure, extension and shear of cuboid

We consider an undeformed cuboid bounded by the planes

$$x_1 = A_1, \quad x_1 = A_2, \quad x_2 = \pm B, \quad x_3 = \pm C. \tag{3.1}$$

We take ϑ^i to be given by (1.1) where

$$y_1 = r\cos\vartheta, \quad y_2 = r\sin\vartheta \tag{3.2}$$

and we consider the deformation

$$r = f(x_1), \quad \vartheta = \frac{x_2}{\lambda k} + \varphi(x_1), \quad y_3 = \lambda\, x_3 + w(x_1), \tag{3.3}$$

where f, φ and w are functions of x_1 and λ, k are constants. Any plane $x_1 =$ constant then becomes a cylindrical surface $r =$ constant and the cuboid becomes a cylinder bounded by surfaces $r = r_1, r_2$ $(r_1 > r_2)$ corresponding respectively to $x_1 = A_1, A_2$. The constant λ represents an extension ratio parallel to the x_3-axis and $\varphi(x_1)$, $w(x_1)$ represents shearing displacements. If φ and w are non-zero the planes $x_2 = \pm B$, $x_3 = \pm C$ become curved surfaces in the deformed body.

If the body is incompressible then

$$f = (2k\,x_1 + K)^{\frac{1}{2}} \tag{3.4}$$

where K is a constant, but φ and w are unrestricted.

With the help of (3.3) the strain components (2.2) are found to be

$$\begin{aligned} 2e_{11} &= f'^2 + f^2\varphi'^2 + w'^2 - 1, & 2e_{12} &= f^2\varphi'/(\lambda k), \\ 2e_{22} &= f^2/(\lambda^2 k^2) - 1, & 2e_{13} &= \lambda w', \\ 2e_{33} &= \lambda^2 - 1, & e_{23} &= 0, \end{aligned} \tag{3.5}$$

and, from (2.4),

$$I_3 = f^2 f'^2 / k^2, \tag{3.6}$$

where primes denote differentiation with respect to x_1.

The stress tensor (2.3) for a compressible body has components

$$\begin{aligned}
\tau^{11} &= \frac{k f'}{f} \frac{\partial W}{\partial e_{11}}, \\
\tau^{22} &= \frac{1}{\lambda^2 k f f'} \frac{\partial W}{\partial e_{22}} + \frac{k \varphi'^2}{f f'} \frac{\partial W}{\partial e_{11}} + \frac{\varphi'}{\lambda f f'} \left(\frac{\partial W}{\partial e_{12}} + \frac{\partial W}{\partial e_{21}} \right), \\
\tau^{33} &= \frac{\lambda^2 k}{f f'} \frac{\partial W}{\partial e_{33}} + \frac{k w'^2}{f f'} \frac{\partial W}{\partial e_{11}} + \frac{\lambda k w'}{f f'} \left(\frac{\partial W}{\partial e_{13}} + \frac{\partial W}{\partial e_{31}} \right), \\
\tau^{12} &= \frac{1}{2 \lambda f} \left(\frac{\partial W}{\partial e_{12}} + \frac{\partial W}{\partial e_{21}} \right) + \frac{k \varphi'}{f} \frac{\partial W}{\partial e_{11}}, \\
\tau^{13} &= \frac{\lambda k}{2 f} \left(\frac{\partial W}{\partial e_{13}} + \frac{\partial W}{\partial e_{31}} \right) + \frac{k w'}{f} \frac{\partial W}{\partial e_{11}}, \\
\tau^{23} &= \frac{1}{2 f f'} \left(\frac{\partial W}{\partial e_{23}} + \frac{\partial W}{\partial e_{32}} \right) + \frac{k \varphi' w'}{f f'} \frac{\partial W}{\partial e_{11}} + \\
&\quad + \frac{k \lambda \varphi'}{2 f f'} \left(\frac{\partial W}{\partial e_{13}} + \frac{\partial W}{\partial e_{31}} \right) + \frac{w'}{2 \lambda f f'} \left(\frac{\partial W}{\partial e_{12}} + \frac{\partial W}{\partial e_{21}} \right).
\end{aligned} \tag{3.7}$$

If the body is incompressible we must add terms p, p/r^2, p to τ^{11}, τ^{22}, τ^{33} respectively in (3.7), and we can make use of (3.4).

In the compressible case the stresses τ^{ij} are functions of x_1 (or r) only. The equations of equilibrium (2.6), in cylindrical polar coordinates (r, ϑ, y_3), reduce to

$$\begin{aligned}
\frac{d \tau^{11}}{d r} + \frac{\tau^{11} - r^2 \tau^{22}}{r} &= 0, \\
\frac{d \tau^{12}}{d r} + \frac{3 \tau^{12}}{r} &= 0, \\
\frac{d \tau^{13}}{d r} + \frac{\tau^{13}}{r} &= 0.
\end{aligned} \tag{3.8}$$

From the second and third of (3.8) we obtain

$$r^3 \tau^{12} = \alpha, \qquad r \tau^{13} = \beta, \tag{3.9}$$

where α and β are constants. If the surfaces $r = r_1$, $r = r_2$ of the deformed cuboid are to be free from applied shear stresses, τ^{12} and τ^{31} vanish on both surfaces. In this case

$$\tau^{12} \equiv 0, \qquad \tau^{13} \equiv 0,$$

and α and β are zero in (3.9). Also, from (3.7), we then have

$$\begin{aligned}
2 k \lambda \varphi' \frac{\partial W}{\partial e_{11}} + \frac{\partial W}{\partial e_{12}} + \frac{\partial W}{\partial e_{21}} &= 0, \\
\frac{2 w'}{\lambda} \frac{\partial W}{\partial e_{11}} + \frac{\partial W}{\partial e_{13}} + \frac{\partial W}{\partial e_{31}} &= 0.
\end{aligned} \tag{3.10}$$

If, in addition, there are no terms of the first power in e_{12} or e_{13} in W then* one solution of (3.10) is

$$\varphi' = w' = 0. \tag{3.11}$$

We then recover the deformation considered by GREEN & ADKINS (1960) and all the formulae they give are correct** if this restriction is imposed on W.

Suppose W is not subject to this restriction. Equations (3.10), are then polynomial equations for the functions φ' and w' in terms of x_1 and $f(x_1)$ and the first of equations (3.8) gives a differential equation for $f(x_1)$. When W is a general polynomial in the components of strain it appears to be very difficult to determine whether or not equations (3.10) yield an appropriate real solution for φ' and w'; there may, of course, be more than one possible set of values of φ' and w' satisfying (3.10). We assume, for the present, that at least one suitable set of real values of φ' and w' in terms of x_1 and $f(x_1)$ can be found to satisfy (3.10), and thus make τ^{12}, τ^{13} vanish identically. The remaining stresses in (3.7) then reduce to

$$\begin{aligned}
\tau^{11} &= \frac{k f'}{f} \frac{\partial W}{\partial e_{11}}, \\
\tau^{22} &= \frac{1}{\lambda^2 k f f'} \frac{\partial W}{\partial e_{22}} - \frac{k \varphi'^2}{f f'} \frac{\partial W}{\partial e_{11}}, \\
\tau^{33} &= \frac{\lambda^2 k}{f f'} \frac{\partial W}{\partial e_{33}} - \frac{k w'^2}{f f'} \frac{\partial W}{\partial e_{11}}, \\
\tau^{23} &= \frac{1}{2 f f'} \left(\frac{\partial W}{\partial e_{23}} + \frac{\partial W}{\partial e_{32}} \right) - \frac{k \varphi' w'}{f f'} \frac{\partial W}{\partial e_{11}}.
\end{aligned} \tag{3.12}$$

Since the components of strain e_{ij} in (3.5) are independent of x_2 and x_3 it follows that W may be regarded as a function of x_1. Then, using (3.5) and (3.10), we have

$$\begin{aligned}
\frac{dW(x_1)}{dx_1} &= \frac{\partial W}{\partial e_{11}} (f' f'' + f f' \varphi'^2 + f^2 \varphi' \varphi'' + w' w'') + \\
&\quad + \frac{\partial W}{\partial e_{22}} \frac{f f'}{\lambda^2 k^2} + \frac{1}{2} \left(\frac{\partial W}{\partial e_{12}} + \frac{\partial W}{\partial e_{21}} \right) \left(\frac{2 f f' \varphi'}{\lambda k} + \frac{f^2 \varphi''}{\lambda k} \right) + \frac{\lambda w''}{2} \left(\frac{\partial W}{\partial e_{13}} + \frac{\partial W}{\partial e_{31}} \right) \\
&= \frac{\partial W}{\partial e_{11}} (f' f'' - f f' \varphi'^2) + \frac{\partial W}{\partial e_{22}} \frac{f f'}{\lambda^2 k^2}
\end{aligned}$$

and with the help of (3.12) and the first of equations (3.8) this further reduces to

$$\frac{dW(x_1)}{dx_1} = \frac{1}{k} \frac{d}{dx_1} (f f' \tau^{11}).$$

This equation can be integrated to give

$$\tau^{11} = \frac{k}{f f'} [W(x_1) + W_0] \tag{3.13}$$

where W_0 is a constant. From (3.13) and the first of equations (3.12) we have

$$f'^2 \frac{\partial W}{\partial e_{11}} = W(x_1) + W_0, \tag{3.14}$$

---

* Terms containing the product $e_{12} e_{13}$ are *not*, however, excluded from W.

** This applies to §§ 2.8, 2.14. Corresponding restrictions on W are needed in the remaining sections 2.9, 2.10, 2.15—2.18 of Chapter II.

a differential equation for $f(x_1)$, assuming that φ', w' have been found in terms of x_1 and $f(x_1)$ from (3.10).

When the body is incompressible the function $f(x_1)$ is defined by (3.4). The functions φ', w' are then determined from (3.10) and, assuming this to be possible, equations (3.13) and (3.14) are then replaced by

$$\tau^{11} = \frac{k^2}{f^2} \frac{\partial W}{\partial e_{11}} + p = W(x_1) + W_0, \tag{3.15}$$

where p is a function of x_1. In addition, the remaining stresses in (3.12) become

$$\begin{aligned} \tau^{22} &= \frac{1}{\lambda^2 k^2} \frac{\partial W}{\partial e_{22}} - \varphi'^2 \frac{\partial W}{\partial e_{11}} + \frac{p}{r^2}, \\ \tau^{33} &= \lambda^2 \frac{\partial W}{\partial e_{33}} - w'^2 \frac{\partial W}{\partial e_{11}} + p, \\ \tau^{23} &= \frac{1}{2k} \left(\frac{\partial W}{\partial e_{23}} + \frac{\partial W}{\partial e_{32}} \right) - \varphi' w' \frac{\partial W}{\partial e_{11}}. \end{aligned} \tag{3.16}$$

If the curved surfaces $r = r_1$, r_2 of the deformed cuboid are to be entirely free from applied stress then, in addition to (3.10), we must have

$$W(A_1) = W(A_2) = -W_0, \tag{3.17}$$

for both compressible and incompressible bodies. Once again, further consideration is necessary before we know whether or not it is possible to satisfy (3.17).

The faces $x_2 = \pm B$ of the cuboid become curved surfaces in the deformed state. In order to find the resultant forces acting on these surfaces it is, perhaps, simplest to consider stresses acting across any plane $\vartheta =$ constant of the deformed body. These are equivalent to a normal force N_2, a shear force S_2 parallel to the y_3-axis, and a couple M_2 about the y_3-axis where

$$\begin{aligned} N_2 &= \iint r^2 \tau^{22} \, dr \, dy_3, \\ S_2 &= \iint r \tau^{23} \, dr \, dy_3, \\ M_2 &= \iint r^3 \tau^{22} \, dr \, dy_3, \end{aligned}$$

the integration being over the section $\vartheta =$ constant of the deformed cuboid. The integrands can be regarded as functions of r and the integrations with respect to y_3 can be performed immediately between the limits $-\lambda C + w(x_1)$ and $\lambda C + w(x_1)$. Then, changing the remaining variable to x_1, we have

$$\begin{aligned} N_2 &= 2\lambda C \int_{A_2}^{A_1} \frac{d}{dx_1} (f \tau^{11}) \, dx_1 \\ &= 2\lambda C \left[f \tau^{11} \right]_{A_2}^{A_1}, \end{aligned} \tag{3.18}$$

where we have used the first of equations (3.8) in the form

$$f' f^2 \tau^{22} = \frac{d}{dx_1} (f \tau^{11}). \tag{3.19}$$

Also

$$\begin{aligned} M_2 &= 2\lambda C \int_{A_2}^{A_1} f \frac{d}{d x_1} (f \tau^{11})\, d x_1 \\ &= 2\lambda C \left\{ [f^2 \tau^{11}]_{A_2}^{A_1} - k \int_{A_2}^{A_1} [W(x_1) + W_0]\, d x_1 \right\}. \end{aligned} \tag{3.20}$$

If equations (3.17) can be satisfied then (3.18) and (3.20) reduce to

$$\begin{aligned} N_2 &= 0, \\ M_2 &= 2\lambda k C \left[\int_{A_1}^{A_2} W(x_1)\, d x_1 - (A_1 - A_2) W_0 \right]. \end{aligned} \tag{3.21}$$

These formulae, together with (3.18) and (3.20), are valid for compressible and incompressible bodies and are similar in form to those given by GREEN & ADKINS*. They *may* be valid for a completely general strain energy function W provided equations (3.10) can be satisfied and, in the case of (3.20), provided also equation (3.17) holds. The doubt about their validity rests on our inability to decide whether or not equations (3.10) have suitable solutions for φ', w'. By introducing the condition that W contains no terms of the first power in e_{12} or e_{13} (apart from $e_{12}\, e_{13}$) but is otherwise unrestricted, we have seen that one solution is $\varphi' \equiv w' \equiv 0$, and this is the case considered by GREEN & ADKINS, although they made some statements which gave a contrary impression.

Forces are required over the curved faces of the cuboid which correspond to the plane faces $y_3 = \pm C$ of the undeformed body, but we omit detailed examination here.

GREEN & ADKINS also considered the flexure of a curvilinearly aeolotropic cuboid and their work requires a similar restriction to be imposed on W**.

---

* Chapter II, § 2.8 and 2.14.

** In their notation W must contain no terms jointly of the first power in $\gamma'_{(12)}$, $\gamma'_{(13)}$.

References

ADKINS, J. E.: Proc. Roy. Soc. A **231**, 75 (1955).
GREEN, A. E., & J. E. ADKINS: Large Elastic Deformations. Oxford 1960.

Department of Mathematics
The University of Nottingham
and
King's College
Newcastle-upon-Tyne

(Received March 15, 1961)

Propagation of a Plane Wave in an Isotropic Elastic Material Subjected to Pure Homogeneous Deformation

M. Hayes & R. S. Rivlin

1. Introduction

In the present paper, we apply the theory [*1*] of the superposition of infinitesimal deformations on finite deformations in an isotropic elastic material to the study of the propagation of a plane wave of small amplitude in an infinite body of the material which is subjected to a static, pure homogeneous deformation. It is seen that the secular equation for the determination of the square of the velocity of propagation in a given direction has three real eigen-values and correspondingly three real mutually perpendicular eigen-directions. Provided these three eigen-values are all positive, travelling waves may be propagated in the body in the direction considered with linear polarisations along each of these eigen-directions. If one or more of the eigen-values is negative for any direction of propagation, the body is inherently unstable in the state of pure homogeneous deformation considered.

The condition for this to occur thus imposes a limitation on the strain-energy function which can be valid for a material that can be maintained in the state of pure, homogeneous deformation considered. In the case of unstressed materials obeying the generalized Hooke's law, this has long since been pointed out by Kelvin [*2*]. It is quite apparent that a similar procedure can be applied in the case of visco-elastic materials, whether linear or non-linear, to obtain restrictions on the constitutive equations.

Finally, it is shown that the necessary conditions for stability of the material which have been derived imply certain inequalities proposed by Baker & Ericksen [*3*], although the converse is not necessarily the case.

2. The fundamental equations

We consider a body of isotropic perfectly-elastic material to be subjected to a finite static deformation, without body forces being applied, in which a particle initially at X_i in the rectangular Cartesian coordinate system x moves to x_i in the same system. We shall call this state of deformation of the body the state A.

The strain-energy function W, per unit volume measured in the undeformed state of the body, is a function of three invariants of the deformation I_1, I_2

and I_3 defined by

$$I_1 = g_{ii},$$

and

$$I_2 = \tfrac{1}{2}(g_{ii} g_{jj} - g_{ij} g_{ji}) \tag{2.1}$$

$$I_3 = |g_{ij}|,$$

where g_{ij} is defined by

$$g_{ij} = \frac{\partial x_i}{\partial X_k} \frac{\partial x_j}{\partial X_k}. \tag{2.2}$$

The components of stress σ_{ij} in the coordinate system x are given [4] by

$$\sigma_{ij} = \frac{2}{I_3^{\frac{1}{2}}} [(W_1 + I_1 W_2) g_{ij} - g_{ik} g_{kj} W_2 + I_3 W_3 \delta_{ij}], \tag{2.3}$$

where W_1, W_2 and W_3 denote $\partial W/\partial I_1$, $\partial W/\partial I_2$ and $\partial W/\partial I_3$ respectively. Since the body is in equilibrium and no body forces are applied,

$$\frac{\partial \sigma_{ij}}{\partial x_j} = 0. \tag{2.4}$$

We now consider that a further deformation, which may vary with time, takes place, in which the particle x_i moves to $\bar{x}_i$ where

$$\bar{x}_i = x_i + \varepsilon u_i. \tag{2.5}$$

ε is a constant, which is small in the sense that terms of second and higher degrees in ε may be neglected in comparison with those of first degree. In the resulting state of deformation the body will be said to be in state B. Let $\sigma_{ij} + \varepsilon \bar{\sigma}_{ij}$ be the components of stress in the coordinate system x, when the body is in state B, at the particle which is at X_i in the undeformed state. It has been shown in the previous paper [1] that

$$\begin{aligned} \bar{\sigma}_{ij} = 2 I_3^{-\frac{1}{2}} \{ & -\tfrac{1}{2} \bar{I}_3 I_3^{-1} [g_{ij}(W_1 + I_1 W_2) - g_{ik} g_{kj} W_2 - I_3 W_3 \delta_{ij}] + \\ & + \bar{g}_{ij}(W_1 + I_1 W_2) + g_{ij}(\bar{W}_1 + \bar{I}_1 W_2 + I_1 \bar{W}_2) - \\ & - g_{ik} g_{kj} \bar{W}_2 - W_2 (g_{ik} \bar{g}_{kj} + \bar{g}_{ik} g_{kj}) + I_3 \bar{W}_3 \delta_{ij} \}, \end{aligned} \tag{2.6}$$

where

$$\bar{g}_{ij} = \frac{\partial x_i}{\partial X_k} \frac{\partial u_j}{\partial X_k} + \frac{\partial x_j}{\partial X_k} \frac{\partial u_i}{\partial X_k}, \tag{2.7}$$

$$\begin{aligned} \bar{I}_1 &= \bar{g}_{ij}, \\ \bar{I}_2 &= g_{ii} \bar{g}_{jj} - g_{ij} \bar{g}_{ji}, \\ \bar{I}_3 &= 2 I_3^{\frac{1}{2}} \varepsilon_{ijk} \left(\frac{\partial u_i}{\partial X_1} \frac{\partial x_j}{\partial X_2} \frac{\partial x_k}{\partial X_3} + \frac{\partial u_i}{\partial X_2} \frac{\partial x_j}{\partial X_3} \frac{\partial x_k}{\partial X_1} + \frac{\partial u_i}{\partial X_3} \frac{\partial x_j}{\partial X_1} \frac{\partial x_k}{\partial X_2} \right), \end{aligned} \tag{2.8}$$

and $\bar{W}_1$, $\bar{W}_2$, $\bar{W}_3$ are defined by

$$\bar{W}_i = W_{ij} \bar{I}_j, \tag{2.9}$$

where

$$W_{ij} = \frac{\partial^2 W}{\partial I_i \partial I_j}. \tag{2.10}$$

The equation of motion for the body in state B is

$$\frac{\partial}{\partial \bar{x}_j} (\sigma_{ij} + \varepsilon \bar{\sigma}_{ij}) = \varepsilon \bar{\varrho} \frac{\partial^2 u_i}{\partial t^2}, \tag{2.11}$$

where $\bar{\varrho}$ is the density of the material in state B. If ϱ is the density of the material in state A, then with the neglect of terms of higher degree than the first in ε, we have

$$\bar{\varrho} = \varrho\left(1 - \varepsilon \frac{\partial u_i}{\partial x_i}\right). \tag{2.12}$$

Introducing (2.12) and (2.5) into (2.11) and neglecting terms of higher degree than the first in ε, we obtain

$$\bar{\sigma}_{ij,j} - \sigma_{ij,k}\, u_{k,j} = \varrho\, \partial^2 u_i / \partial t^2. \tag{2.13}$$

The notation $_{,i}$ is used to denote $\partial/\partial x_i$. In the case when the state of deformation A is a pure homogeneous deformation, $\sigma_{ij,k} = 0$ and equation (2.13) becomes

$$\bar{\sigma}_{ij,j} = \varrho\, \partial^2 u_i / \partial t^2. \tag{2.14}$$

3. Small deformation superposed on pure homogeneous deformation

We now consider the case when the state of deformation A is a pure homogeneous deformation with extension ratios λ_α ($\alpha = 1, 2, 3$) and principal directions parallel to the axes of the coordinate system x. The deformation is described by*

$$x_\alpha = \lambda_\alpha X_\alpha. \tag{3.1}$$

Introducing (3.1) into (2.2), we have

$$g_{\alpha\beta} = \lambda_\alpha \lambda_\beta \delta_{\alpha\beta} = \lambda_\alpha^2 \delta_{\alpha\beta}. \tag{3.2}$$

With (2.2), we obtain

$$\begin{aligned} I_1 &= \sum_{\alpha=1}^{3} \lambda_\alpha^2, \\ I_2 &= \tfrac{1}{2} \sum_{\substack{\alpha,\beta=1 \\ \alpha\neq\beta}}^{3} \lambda_\alpha^2 \lambda_\beta^2, \\ I_3 &= \lambda_1^2 \lambda_2^2 \lambda_3^2. \end{aligned} \tag{3.3}$$

Introducing (3.1) into (2.7), we obtain

$$\begin{aligned} \bar{g}_{\alpha\beta} &= \lambda_\alpha \frac{\partial u_\beta}{\partial X_\alpha} + \lambda_\beta \frac{\partial u_\alpha}{\partial X_\beta} \\ &= \lambda_\alpha^2 u_{\beta,\alpha} + \lambda_\beta^2 u_{\alpha,\beta}. \end{aligned} \tag{3.4}$$

Introducing (3.4) and (3.1) into (2.8)

$$\begin{aligned} \bar{I}_1 &= \sum_{\alpha=1}^{3} 2\lambda_\alpha^2 u_{\alpha,\alpha}, \\ \bar{I}_2 &= 2 \sum_{\alpha=1}^{3} \lambda_\alpha^2 (I_1 - \lambda_\alpha^2)\, u_{\alpha,\alpha}, \\ \bar{I}_3 &= 2 I_3 \sum_{\alpha=1}^{3} u_{\alpha,\alpha}. \end{aligned} \tag{3.5}$$

---

* The summation convention will not apply when Greek subscripts are used.

Introducing (3.2), (3.4) and (3.5) into (2.6), we obtain

$$\begin{aligned}\bar{\sigma}_{\alpha\beta} = 2I_3^{-\frac{1}{2}}\Big\{&(W_1 + I_1 W_2)\,[-u_{k,k}\,\lambda_\alpha^2\,\delta_{\alpha\beta} + \lambda_\alpha^2\,u_{\beta,\alpha} + \lambda_\beta^2\,u_{\alpha,\beta}] + \\ &+ W_2\Big[\lambda_\alpha^2\,\delta_{\alpha\beta}\Big(\lambda_\alpha^2\,u_{k,k} + 2\sum_{\gamma=1}^{3}\lambda_\gamma^2\,u_{\gamma,\gamma}\Big) - (\lambda_\alpha^2 + \lambda_\beta^2)\,(\lambda_\alpha^2\,u_{\beta,\alpha} + \lambda_\beta^2\,u_{\alpha,\beta})\Big] + \\ &+ W_3\,I_3\,u_{k,k}\,\delta_{\alpha\beta} + [\lambda_\alpha^2(\overline{W}_1 + I_1\,\overline{W}_2 - \lambda_\alpha^2\,\overline{W}_2) + I_3\,\overline{W}_3]\,\delta_{\alpha\beta}\Big\}.\end{aligned} \tag{3.6}$$

Thus

$$\bar{\sigma}_{\alpha\beta} = 2I_3^{-\frac{1}{2}}\{W_1 + (I_1 - \lambda_\alpha^2 - \lambda_\beta^2)\,W_2\}\,(\lambda_\alpha^2\,u_{\beta,\alpha} + \lambda_\beta^2\,u_{\alpha,\beta}), \qquad \alpha \neq \beta,$$

and

$$\begin{aligned}\bar{\sigma}_{\alpha\alpha} = 2I_3^{-\frac{1}{2}}\Big\{&(W_1 + I_1 W_2)\,\lambda_\alpha^2(2u_{\alpha,\alpha} - u_{k,k}) + \\ &+ W_2\,\lambda_\alpha^2\Big[\lambda_\alpha^2(u_{k,k} - 4u_{\alpha,\alpha}) + 2\sum_{\gamma=1}^{3}\lambda_\gamma^2\,u_{\gamma,\gamma}\Big] + \\ &+ W_3\,I_3\,u_{k,k} + \\ &+ [\lambda_\alpha^2(\overline{W}_1 + I_1\,\overline{W}_2 - \lambda_\alpha^2\,\overline{W}_2) + I_3\,\overline{W}_3]\Big\}.\end{aligned} \tag{3.7}$$

We note, from (2.9), that

$$\begin{aligned}\lambda_\alpha^2(\overline{W}_1 + I_1\,\overline{W}_2 - \lambda_\alpha^2\,\overline{W}_2) + I_3\,\overline{W}_3 = &\{\lambda_\alpha^2[W_{11} + (I_1 - \lambda_\alpha^2)\,W_{21}] + I_3\,W_{31}\}\,\bar{I}_1 + \\ &+ \{\lambda_\alpha^2[W_{12} + (I_1 - \lambda_\alpha^2)\,W_{22}] + I_3\,W_{32}\}\,\bar{I}_2 + \\ &+ \{\lambda_\alpha^2[W_{13} + (I_1 - \lambda_\alpha^2)\,W_{23}] + I_3\,W_{33}\}\,\bar{I}_3.\end{aligned} \tag{3.8}$$

Introducing (3.5) into (3.8) and the relation so obtained into the second of equations (3.7), we see that $\bar{\sigma}_{\alpha\alpha}$ may be written in the form

$$\bar{\sigma}_{\alpha\alpha} = \sum_{\beta=1}^{3} c_{\alpha\beta}\,u_{\beta,\beta}, \tag{3.9}$$

where $c_{\alpha\beta}$ is given by

$$\begin{aligned}c_{\alpha\beta} = 2I_3^{-\frac{1}{2}}\{&[W_1 + W_2(I_1 - \lambda_\alpha^2 - \lambda_\beta^2)]\,\lambda_\alpha^2(2\delta_{\alpha\beta} - 1) + \lambda_\alpha^2\,\lambda_\beta^2\,W_2 + I_3\,W_3 + \\ &+ 2\{\lambda_\alpha^2[W_{11} + (I_1 - \lambda_\alpha^2)\,W_{21}] + I_3\,W_{31}\}\,\lambda_\beta^2 + \\ &+ 2\{\lambda_\alpha^2[W_{12} + (I_1 - \lambda_\alpha^2)\,W_{22}] + I_3\,W_{32}\}\,\lambda_\beta^2(I_1 - \lambda_\beta^2) + \\ &+ 2\{\lambda_\alpha^2[W_{13} + (I_1 - \lambda_\alpha^2)\,W_{23}] + I_3\,W_{33}\}\,I_3\}.\end{aligned} \tag{3.10}$$

4. Propagation of plane waves

We now assume that the displacements εu_i arise from the propagation through the deformed material of a plane wave, with wave number k, in a direction with direction-cosines l_i in the coordinate system x and that u_i is given by

$$u_i = U_i(t)\,e^{\iota k z}, \tag{4.1}$$

where

$$z_i = l_i\,x_i, \qquad l_i\,l_i = 1. \tag{4.2}$$

Then, z is the distance measured along the direction with direction-cosines l_i.

By introducing (4.1) and (4.2) into the first of equations (3.7), we obtain

$$\bar{\sigma}_{\alpha\beta} = 2 I_3^{-\frac{1}{2}} \{W_1 + (I_1 - \lambda_\alpha^2 - \lambda_\beta^2) W_2\} (\lambda_\alpha^2 l_\alpha U_\beta + \lambda_\beta^2 l_\beta U_\alpha)\, \iota k e^{\iota k z} \qquad (\alpha \neq \beta). \tag{4.3}$$

Similarly, by introducing (4.1) and (4.2) into equation (3.9), we obtain

$$\bar{\sigma}_{\alpha\alpha} = \sum_{\gamma=1}^{3} c_{\alpha\gamma} l_\gamma U_\gamma\, \iota k e^{\iota k z}. \tag{4.4}$$

Introducing (4.3) and (4.4) into (2.14), we obtain an equation of the form

$$\sum_{\beta=1}^{3} (k^2 s_{\alpha\beta} + \tfrac{1}{2} I_3^{\frac{1}{2}} \varrho\, \delta_{\alpha\beta}\, \partial^2/\partial t^2)\, U_\beta = 0, \tag{4.5}$$

where

$$\begin{aligned} s_{\alpha\beta} = {} & \lambda_\alpha^2 l_\alpha l_\beta (1 - 2\delta_{\alpha\beta}) [W_1 + W_2 (I_1 - \lambda_\alpha^2 - \lambda_\beta^2)] + \\ & + \delta_{\alpha\beta} \left\{ \sum_{\gamma=1}^{3} \lambda_\gamma^2 l_\gamma^2 [W_1 + W_2 (I_1 - \lambda_\alpha^2 - \lambda_\gamma^2)] \right\} + \tfrac{1}{2} I_3^{\frac{1}{2}} c_{\alpha\beta} l_\alpha l_\beta. \end{aligned} \tag{4.6}$$

Writing

$$c_{\alpha\beta} = \bar{c}_{\alpha\beta} + 2 I_3^{-\frac{1}{2}} \{[W_1 + W_2 (I_1 - \lambda_\alpha^2 - \lambda_\beta^2)]\, \lambda_\alpha^2 (2\delta_{\alpha\beta} - 1) + \lambda_\alpha^2 \lambda_\beta^2 W_2 + I_3 W_3\}, \tag{4.7}$$

we obtain from (4.6)

$$\begin{aligned} s_{\alpha\beta} = {} & (\lambda_\alpha^2 \lambda_\beta^2 W_2 + I_3 W_3)\, l_\alpha l_\beta + \\ & + \delta_{\alpha\beta} \left\{ \sum_{\gamma=1}^{3} \lambda_\gamma^2 l_\gamma^2 [W_1 + W_2 (I_1 - \lambda_\alpha^2 - \lambda_\gamma^2)] \right\} + \tfrac{1}{2} I_3^{\frac{1}{2}} \bar{c}_{\alpha\beta} l_\alpha l_\beta. \end{aligned} \tag{4.8}$$

Replacing $\partial/\partial t$ by $\iota\omega$ in (4.5), the secular equation may be written as

$$|s_{\alpha\beta} - C^2 \delta_{\alpha\beta}| = 0, \tag{4.9}$$

where

$$C^2 = \tfrac{1}{2} I_3^{\frac{1}{2}} \varrho\, \omega^2/k^2. \tag{4.10}$$

It follows from (3.10) and (4.7) that $\bar{c}_{\alpha\beta}$ is symmetric. Hence from (4.8) $s_{\alpha\beta}$ is also symmetric.

5. Discussion of the results

Equations (4.5) and (4.9) can be transformed to any other rectangular Cartesian coordinate system y, related to the system x by the transformation

$$y_i = a_{ij} x_j, \tag{5.1}$$

with

$$a_{ik} a_{jk} = a_{ki} a_{kj} = \delta_{ij}, \qquad |a_{ij}| = 1. \tag{5.2}$$

Let $\varepsilon V_i(t)\, e^{\iota k z}$ be the displacement components in the system y. Then

$$V_i = a_{ij} U_j. \tag{5.3}$$

We note that z is unaltered by the transformation from the system x to the system y. With the relation (5.3), it is evident that equation (4.5) can be expressed in the form

$$(k^2 t_{ij} + \tfrac{1}{2} I_3^{\frac{1}{2}} \varrho\, \delta_{ij}\, \partial^2/\partial t^2)\, V_j = 0, \tag{5.4}$$

where

$$t_{ij} = a_{ip} a_{jq} s_{pq}. \tag{5.5}$$

With the same notation, the secular equation (4.9) can be written in the form

$$|t_{ij} - C^2 \delta_{ij}| = 0. \tag{5.6}$$

Since $\|s_{\alpha\beta}\|$ is symmetric, the coordinate system y can be chosen in such a way that $t_{ij}=0$ $(i \neq j)$. Then, it is seen that equation (5.4) takes the form

$$\left(t_{\alpha\alpha} + \frac{\frac{1}{2} I_3^{\frac{1}{3}} \varrho}{k^2} \frac{\partial^2}{\partial t^2}\right) V_\alpha = 0. \tag{5.7}$$

Thus, the directions of the axes of y are the three directions of polarisation for which waves, which are not mutually coupled, may be propagated in the direction of z. The velocity of propagation, along the direction z, of the wave polarised in the direction of y_α is $(t_{\alpha\alpha}/\frac{1}{2} I_3^{\frac{1}{3}} \varrho)^{\frac{1}{2}}$. $t_{\alpha\alpha}$ $(\alpha = 1, 2, 3)$, are, of course, the solutions for C^2 of the secular equation (4.9). Provided all of these solutions are positive, the velocities of propagation of all three waves are real.

However, if one or more of the solutions of (4.9) for C^2 is negative for any direction of propagation of the plane wave, *i.e.* for any real values of l_i satisfying the relation $l_i l_i = 1$, it is easily shown that the material is inherently unstable in the state of pure homogeneous deformation considered. Equation (4.5) may still be written in the form (5.7). Let us suppose that t_{11} is negative, then the general solution of (5.7) with $\alpha = 1$ is

$$V_1 = A_1 e^{\omega_1 t} + B_1 e^{-\omega_1 t}, \tag{5.8}$$

where A_1 and B_1 are integration constants and

$$\omega_1 = k \left(-\frac{t_{11}}{\frac{1}{2} I_3^{\frac{1}{3}} \varrho}\right)^{\frac{1}{2}}. \tag{5.9}$$

The displacement component $\varepsilon V_1(t)\, e^{\iota k z}$ is then given by

$$\varepsilon V_1(t)\, e^{\iota k z} = \varepsilon A_1 e^{\omega_1 t + \iota k z} + \varepsilon B_1 e^{-\omega_1 t + \iota k z}, \tag{5.10}$$

We shall now assume that when $t=0$ the rate of change of the displacement is zero. Then, from (5.10), $A_1 = B_1$ and we can re-write (5.10) as

$$\varepsilon V_1(t)\, e^{\iota k z} = \varepsilon A_1 \left(e^{\omega_1 t + \iota k z} + e^{-\omega_1 t + \iota k z}\right). \tag{5.11}$$

It is apparent that a displacement parallel to the axis y_1, which is sinusoidal along the direction z with wave number k, builds up exponentially with time.

We thus have the result that if one or more of the values of C^2 given by (4.9) is negative, for any direction of propagation of the plane wave, the material considered is inherently unstable in the state of pure homogeneous deformation described by (3.1).

For a real material which can be maintained in the state of pure homogeneous deformation (3.1), the dependence of W on I_1, I_2 and I_3 must therefore be such that the three roots of (4.9) for C^2 are all positive, for all l_i in the expression (4.6) which satisfy $l_i l_i = 1$. The conditions for this to be the case may be written in a variety of forms. For example, a necessary and sufficient condition that the roots of (4.9) be positive is that the quadratic form $s_{ij} y_i y_j$ be positive definite. Alternatively, the necessary and sufficient conditions may be written

$$\operatorname{tr} \boldsymbol{S} > 0, \qquad (\operatorname{tr} \boldsymbol{S})^2 - \operatorname{tr} \boldsymbol{S}^2 > 0 \quad \text{and} \quad |\boldsymbol{S}| > 0, \tag{5.12}$$

where the notation $\boldsymbol{S} = \|s_{\alpha\beta}\|$ is used.

We note that the conditions (5.12) are independent of frequency. Consequently, if they are violated for some l_i and for some frequency, or wave-length, they will be violated for all frequencies and wave-lengths. If the conditions are not satisfied for some values of I_1, I_2 and I_3 and for some real value of l_i, satisfying $l_i l_i = 1$, any infinitesimal vibration present, due to Brownian motion or other cause, will build up. The material cannot then be held in the state of pure, homogeneous deformation considered. Conversely, if a material can be held in the state of pure, homogeneous deformation considered, then a strain-energy function which violates the condition (5.12) for this state of deformation and some value of l is inappropriate to the description of the properties of the material.

6. Relation to the Stability Criteria of Baker and Ericksen

BAKER & ERICKSEN [*3*] proposed as a necessary restriction on the strain-energy function for an isotropic elastic material, that, when subjected to a pure homogeneous deformation with principal extension ratios λ_α, the relation

$$W_1 + \lambda_\alpha^2 W_2 \geqq 0 \tag{6.1}$$

must be satisfied. It is easily seen that the relations (5.12) imply the relation (6.1), while the converse is not the case.

We may re-write (4.8) as

$$s_{\alpha\beta} = \delta_{\alpha\beta} \sum_{\gamma=1}^{3} A_{\alpha\gamma} l_\gamma^2 + (1 - \delta_{\alpha\beta}) B_{\alpha\beta} l_\alpha l_\beta, \tag{6.2}$$

where

and

$$\begin{aligned} A_{\alpha\beta} &= \delta_{\alpha\beta}(\lambda_\beta^4 W_2 + I_3 W_2 + \tfrac{1}{2} I_3^{\frac{1}{3}} \bar{c}_{\alpha\beta}) + \lambda_\beta^2 \{W_1 + (I_1 - \lambda_\alpha^2 - \lambda_\beta^2) W_2\} \\ B_{\alpha\beta} &= \lambda_\alpha^2 \lambda_\beta^2 W_2 + I_3 W_3 + \tfrac{1}{2} I_3^{\frac{1}{3}} \bar{c}_{\alpha\beta} \qquad (\alpha \neq \beta). \end{aligned} \tag{6.3}$$

Consequently,

and

$$\begin{aligned} \operatorname{tr} \boldsymbol{S} &= \sum_{\alpha=1}^{3} s_{\alpha\alpha} = \sum_{\alpha=1}^{3} \Big(\sum_{\beta=1}^{3} A_{\alpha\beta} l_\beta^2 \Big) = \sum_{\beta=1}^{3} l_\beta^2 \Big(\sum_{\alpha=1}^{3} A_{\alpha\beta} \Big) \\ \operatorname{tr} \boldsymbol{S}^2 &= \sum_{\gamma,\eta=1}^{3} \Big\{ \sum_{\alpha=1}^{3} A_{\alpha\gamma} A_{\alpha\eta} + (1 - \delta_{\gamma\eta}) B_{\gamma\eta}^2 \Big\} l_\gamma^2 l_\eta^2. \end{aligned} \tag{6.4}$$

From (6.4), we obtain

$$(\operatorname{tr} \boldsymbol{S})^2 - \operatorname{tr} \boldsymbol{S}^2 = \sum_{\gamma,\eta=1}^{3} U_{\gamma\eta} l_\gamma^2 l_\eta^2, \tag{6.5}$$

where

$$U_{\gamma\eta} = \sum_{\alpha,\beta=1}^{3} (A_{\alpha\gamma} A_{\beta\eta} - A_{\alpha\gamma} A_{\alpha\eta}) - (1 - \delta_{\gamma\eta}) B_{\gamma\eta}^2. \tag{6.6}$$

The conditions (5.12) may therefore be written as

and

$$\begin{gathered} \sum_{\beta=1}^{3} l_\beta^2 \Big(\sum_{\alpha=1}^{3} A_{\alpha\beta} \Big) > 0, \\ \sum_{\gamma,\eta=1}^{3} U_{\gamma\eta} l_\gamma^2 l_\eta^2 > 0 \\ |s_{\alpha\beta}| > 0. \end{gathered} \tag{6.7}$$

These inequalities must be valid for all values of l_1^2, l_2^2, l_3^2 which satisfy the relation

$$\sum_{\alpha=1}^{3} l_\alpha^2 = 1. \tag{6.8}$$

Necessary and sufficient conditions that the first of the relations (6.7) be valid are

$$\sum_{\alpha=1}^{3} A_{\alpha\beta} > 0. \tag{6.9}$$

A necessary condition that the second of the relations (6.7) be valid is that

$$U_{\alpha\alpha} > 0. \tag{6.10}$$

With (6.6), the relation (6.10) may be written as

$$\sum_{\substack{\beta,\gamma=1 \\ \beta\neq\gamma}}^{3} A_{\beta\alpha} A_{\gamma\alpha} > 0. \tag{6.11}$$

Finally, a necessary condition that the third of the relations (6.7) be valid is

$$A_{1\alpha} A_{2\alpha} A_{3\alpha} > 0. \tag{6.12}$$

From (6.9), (6.11) and (6.12), it can easily be deduced that

$$A_{\alpha\beta} > 0. \tag{6.13}$$

The inequality (6.1) derived by BAKER & ERICKSEN is a special case of (6.13). For, from (6.13) and (6.3), we have, if $\alpha \neq \beta$,

$$W_1 + (I_1 - \lambda_\alpha^2 - \lambda_\beta^2) W_2 > 0. \tag{6.14}$$

Bearing in mind the definition of I_1, it is evident that (6.14) implies (6.1).

It is apparent that, in general, the relations (6.1) do not imply the relations (5.12), since the latter relations involve the second derivatives of W with respect to the strain invariants I_1, I_2, I_3, while the former do not.

Acknowledgement. The results presented in this paper were obtained in the course of research sponsored by the Office of Naval Research under Contract Nonr-562(10) with Brown University and are, to a great extent, those reported in Brown University Technical Report No. 62 (May 1960) under that contract. The authors gratefully acknowledge valuable discussions with Professors E. STERNBERG and R. T. SHIELD.

References

[1] GREEN, A. E., R. S. RIVLIN & R. T. SHIELD: Proc. Roy. Soc. A **211**, 128 (1952).
[2] LORD KELVIN: Phil. Mag. **26**, 414 (1888).
[3] BAKER, M., & J. L. ERICKSEN: J. Wash. Acad. Sci. **44**, 33 (1954).
[4] RIVLIN, R. S.: Phil. Trans. A **241**, 379 (1948).

Division of Applied Mathematics
Brown University
Providence, Rhode Island

(Received February 26, 1961)

On the Motion of Two Linked Bodies

Č. Masaitis

Communicated by E. Leimanis

1. Introduction

The use of satellites for weather observation, communication, navigation, *etc.*, requires a certain orientation of these bodies in space. Various devices have been proposed for the purpose of damping the nutation of a force-free body and thus stabilizing its orientation. Two examples of such devices are discussed in [*1*] and [*2*]. However, the complexity of the equations and the limitations on the range of the variables make the complete analysis of their motion difficult. Furthermore, the shifting of the center of mass within these systems has obvious engineering disadvantages.

A system of two linked bodies, proposed by S. Zaroodny, seems to offer the same possibilities as the two systems mentioned above but is free of their disadvantages. In [*3*] he and J. W. Bradley have exhibited some numerical solutions which show that with very large friction of a viscous type, *i.e.*, proportional to the relative speed of rotation of two bodies, the motion is of the form of a perturbation on the free nutation. Moderate values of the coefficient of friction produce a damping of nutation or metastable forms of nutation, depending on the initial conditions of motion. Very small friction produces rather complicated motion with markedly different position of points of equilibrium from the case of rigid body. In some cases of the motion in the vicinity of the points of equilibrium it was difficult to distinguish between the accumulation of numerical errors and the metastable character of the motion. Therefore, it was desirable to attempt the analytic treatment of the problem.

Although no elementary solution can be constructed, the equations of motion are sufficiently simple to offer possibilities for a perturbation treatment. The unperturbed system can be chosen as the one without friction, which admits a solution in closed form. In this paper such a solution is constructed and the nature of the equilibrium points investigated. The results could be used for selecting special configurations of a case with friction, for investigation by numerical solution.

2. The Equations of Motion and the First Three Integrals

The system to be considered consists in two rigid bodies B_1 and B_2 mounted on a weightless rigid axle A passing through the centers of mass of B_1 and B_2. Let O be the center of mass of the system, I_1, I_2, I_3 be the principal moments of inertia of B_1 about O, not all of them equal, say

$$I_1 < I_2. \tag{1}$$

Let the principal moments of inertia of B_2 about O be I, I, and I_4. Suppose that the axes of the moments I_3 and I_4 coincide with the axle A, and that B_1 and B_2 can rotate about A without friction. Let the coordinate axes (1, 2, 3) (Fig. 1) through O be parallel to the axes of the moments I_1, I_2 and I_3, respectively. Suppose that no external forces or moments act on the system. Then O can be chosen as the origin of an inertial coordinate system (x, y, z). If $\overline{\omega} = (\omega_1, \omega_2, \omega_3)$ is the angular velocity of B_1, where $\omega_1, \omega_2, \omega_3$ are components along the axes (1, 2, 3), $\overline{M}$ the moment of force acting on B_1, and $\overline{H}_1 = (I_1\omega_1, I_2\omega_2, I_3\omega_3)$ the angular momentum of B_1 about 0, then, since O is the origin of an inertial coordinate system, it follows that $\frac{d}{dt}\overline{H}_1 = \overline{M}$, where $\frac{d}{dt} = \frac{d}{d\tau} + \overline{\omega}\times$ and $\frac{d}{dt}$ is the derivative relative to inertial axes (x, y, z), $\frac{d}{d\tau}$ is that relative to rotating axes (1, 2, 3), and $\overline{\omega}$ is the angular velocity of the latter. Therefore, the equations of motion of B_1 are

$$\begin{aligned} &I_1\dot{\omega}_1 + (I_3 - I_2)\,\omega_2\omega_3 = M_1,\\ (2)\qquad &I_2\dot{\omega}_2 + (I_1 - I_3)\,\omega_3\omega_1 = M_2,\\ &I_3\dot{\omega}_3 + (I_2 - I_1)\,\omega_1\omega_2 = M_3. \end{aligned}$$

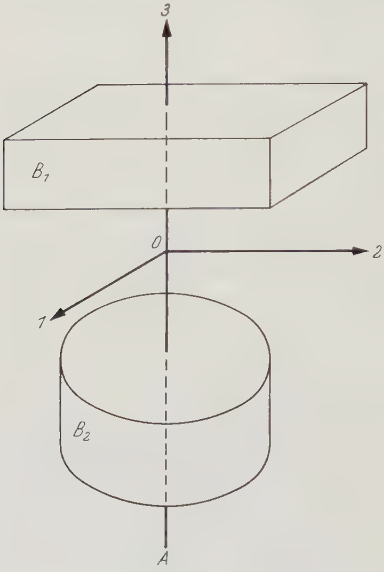

Fig. 1

Because of the assumed axial symmetry of B_2 the angular momentum of B_2 about O is $\overline{H}_2 = (I\omega_1', I\omega_2', I_4\omega_3')$, the components being taken along the axes (1, 2, 3). Let the moment of force about O acting on B_2 be $\overline{M}' = (M_1', M_2', M_3')$. Since B_2 is constrained to rotate on the axle A, there follows $\omega_1' = \omega_1$, $\omega_2' = \omega_2$. Let $\omega_3' = \omega_4$. Applying the operator d/dt defined above to $\overline{H}_2$ yields the equations of motion of B_2:

$$\begin{aligned} &I\dot{\omega}_1 + I_4\omega_2\omega_4 - I\omega_3\omega_2 = M_1',\\ (3)\qquad &I\dot{\omega}_2 - I_4\omega_1\omega_4 + I\omega_3\omega_1 = M_2',\\ &I_4\dot{\omega}_4 \qquad\qquad\qquad\qquad\; = M_3'. \end{aligned}$$

Since no external forces or moments are applied to the system, one has $M_1 = -M_1'$, $M_2 = -M_2'$. Furthermore, since both bodies can rotate independently about axis 3 without interaction, it follows that $M_3 = M_3' = 0$. Hence (2) and (3) yield

$$\begin{aligned} &(I_1 + I)\,\dot{\omega}_1 = \omega_2[(I_2 - I_3 + I)\,\omega_3 - I_4\omega_4],\\ (4)\qquad &(I_2 + I)\,\dot{\omega}_2 = \omega_1[(I_3 - I_1 - I)\,\omega_3 + I_4\omega_4],\\ &I_3\dot{\omega}_3 = \omega_1\omega_2(I_1 - I_2),\\ &I_4\dot{\omega}_4 = 0. \end{aligned}$$

Clearly $I_4\omega_4 = C$, a constant. Let the direction of the axis 3 be so chosen that $C > 0$.

Let

$$\begin{aligned} a_1 &= \frac{I_2 - I_3 + I}{I_1 + I}, \\ a_2 &= \frac{I_3 - I_1 - I}{I_2 + I}, \\ a_3 &= \frac{I_2 - I_1}{I_3} > 0, \\ b_1 &= \frac{C}{I_1 + I} > 0, \\ b_2 &= \frac{C}{I_2 + I} > 0. \end{aligned}$$

By substitution of this in (4) one gets

$$\begin{aligned} \dot{\omega}_1 &= \omega_2 (a_1 \omega_3 - b_1), \\ \dot{\omega}_2 &= \omega_1 (a_2 \omega_3 + b_2), \\ \dot{\omega}_3 &= - a_3 \omega_1 \omega_2. \end{aligned} \tag{5}$$

Multiply the equations (5) by $a_3 b_2 \omega_1$, $a_3 b_1 \omega_2$, $(a_1 b_2 + a_2 b_1) \omega_3$, respectively, add, and integrate. This yields

$$a_3 b_2 \omega_1^2 + a_3 b_1 \omega_2^2 + \beta \omega_3^2 = \beta p^2, \tag{6}$$

where

$$\beta = a_1 b_2 + a_2 b_1 > 0,$$

and p^2 is a constant of integration. Similarly, multiplication of (5) by $a_2 a_3 \omega_1$, $-a_1 a_3 \omega_2$, $-\beta$, respectively, followed by addition and integration yields

$$a_2 a_3 \omega_1^2 - a_1 a_3 \omega_2^2 - 2\beta \omega_3 = 2\beta q, \tag{7}$$

where q is a constant of integration.

3. Singular Points

In the examination of singular points, *i.e.*, points of equilibrium of (5), it is convenient to consider the following cases:

(a) $I < I_3 - I_2$, i.e., $a_1 < 0$, $a_2 > 0$;
(b) $I = I_3 - I_2$, i.e., $a_1 = 0$, $a_2 > 0$;
(c) $I_3 - I_2 < I < I_3 - I_1$, i.e., $a_1 > 0$, $a_2 > 0$;
(d) $I = I_3 - I_1$, i.e., $a_1 > 0$, $a_2 = 0$;
(e) $I > I_3 - I_1$, i.e., $a_1 > 0$, $a_2 < 0$.

In cases (a), (c), and (e) the sets of singular points of (5) are defined by

$$\omega_2 = 0, \quad \omega_3 = -\frac{b_2}{a_2}; \tag{i}$$

$$\omega_1 = 0, \quad \omega_3 = \frac{b_1}{a_1}; \tag{ii}$$

$$\omega_1 = \omega_2 = 0. \tag{iii}$$

In case (b) the set of singular points is defined by (i) and (iii) above, in case (d) by (ii) and (iii). It is of interest to examine which singular points in the phase space $(\omega_1, \omega_2, \omega_3)$ are points of stable equilibrium and which are not. Let $U(P, \varepsilon)$ denote a spherical neighborhood of the point P with radius ε. It is said that the point of equilibrium* P is stable if for every $\varepsilon > 0$ there exists a $\delta(\varepsilon) > 0$ such that if the motion starts at any point in $U(P, \delta)$, it will remain in $U(P, \varepsilon)$ for all values of the independent variable t.

The solution of (6) and (7) for $a_3\omega_1^2$ and $a_3\omega_2^2$ in terms of ω_3 yields

$$a_3\omega_1^2 = a_1(p^2 - \omega_3^2) + 2b_1 q + 2b_1\omega_3, \tag{8}$$

$$a_3\omega_2^2 = a_2(p^2 - \omega_3^2) - 2b_2 q - 2b_2\omega_3. \tag{9}$$

Lemma. *Suppose that p and q are so chosen that* (8) *and* (9) *are satisfied by the coordinates of a point of equilibrium P. If with these values of p and q* (8) *and* (9) *define a non-degenerate curve passing through P, then P is a point of unstable equilibrium.*

The possible types of equilibrium points of (5) are: (1) *nodal points,* (2) *focal points,* (3) *centers,* (4) *saddle points.* However, any solution passing through a singular point P is an intersection of S_1 and S_2, the quadric surfaces defined by (8) and (9), respectively. The intersection has at most four branches passing through P. Hence P cannot be a nodal or a focal point. If $P(\tilde\omega_1, \tilde\omega_2, \tilde\omega_3)$ were a center, then $\omega_1 = \tilde\omega_1$, $\omega_2 = \tilde\omega_2$, $\omega_3 = \tilde\omega_3$ would be a solution of (5) consisting of a single point, *i.e.*, a degenerate curve. Hence the hypothesis of the lemma implies that P is a saddle point, and hence it is unstable, *q.e.d.*

Case (a). *Points of unstable equilibrium.* Let A be an arbitrary point on the line (ii), say, $A\left(0, \tilde\omega_2, \frac{b_1}{a_1}\right)$. In what follows the constants of integration corresponding to initial conditions $(\omega_1, \omega_2, \omega_3)$, $(\tilde\omega_1, \tilde\omega_2, \tilde\omega_3)$, *etc.*, will be denoted by (p, q), $(\tilde p, \tilde q)$, *etc.*, respectively. Since only the second power of p appears in (6), (8), and (9) it will be assumed that $p \geqq 0$. By substituting on $\left(0, \tilde\omega_2, \frac{b_1}{a_1}\right)$ in (6) and (7) it follows that $\tilde q = -\frac{a_1}{2b_1}\left(\tilde p^2 + \frac{b_1^2}{a_1^2}\right)$. With this value of $\tilde q$ formulas (8) and (9) can be written as

$$a_3\omega_1^2 = -a_1\left(\omega_3 - \frac{b_1}{a_1}\right)^2, \tag{10}$$

$$a_3\omega_2^2 = -a_2\omega_3^2 - 2b_2\omega_3 + \frac{\beta}{b_1}\tilde p^2 + \frac{b_1 b_2}{a_1}. \tag{11}$$

If D is one fourth of the discriminant of the right-hand side of (11), then $D = \frac{\beta a_2}{b_1}\left(\tilde p^2 + \frac{b_1 b_2}{a_1 a_2}\right)$. Since in this case $a_1 < 0$ and $\beta = a_1 b_2 + a_2 b_1 > 0$, it follows that $-\frac{b_1}{a_1} > \frac{b_2}{a_2}$ and hence $\frac{b_1^2}{a_1^2} > -\frac{b_1 b_2}{a_1 a_2}$. Since $\tilde\omega_3 = \frac{b_1}{a_1}$ (6) implies that $\tilde p^2 \geqq \frac{b_1^2}{a_1^2}$, hence $\tilde p^2 + \frac{b_1 b_2}{a_1 a_2} > 0$, i.e., $D > 0$. Therefore ω_2 as defined by (11) is real, provided that

---

* Stability in LYAPUNOV's sense.

$$\frac{-b_2-\sqrt{D}}{a_2}\leqq\omega_3\leqq\frac{-b_2+\sqrt{D}}{a_2}. \tag{12}$$

Now $\frac{\beta}{a_1^2}=\frac{a_2}{b_1}\left(\frac{b_1^2}{a_1^2}+\frac{b_1 b_2}{a_1 a_2}\right)$; therefore $\frac{\beta}{a_1^2}\leqq\left(\tilde{p}^2+\frac{b_1 b_2}{a_1 a_2}\right)$ or $\frac{\beta^2}{a_1^2}\leqq D$, whence $-\sqrt{D}\leqq\frac{\beta}{a_1}\leqq\sqrt{D}$, i.e., $-\sqrt{D}\leqq b_2+\frac{a_2 b_1}{a_1}\leqq\sqrt{D}$. Therefore $\frac{-b_2-\sqrt{D}}{a_2}\leqq\frac{b_1}{a_1}\leqq\frac{-b_2+\sqrt{D}}{a_2}$. Since ω_1 as defined by (10) is real for all values of ω_3, this last inequality together with (12) shows that A lies on a non-degenerate integral curve of (5). Hence by the lemma, A is a point of unstable equilibrium, *i.e., the line* (ii) *consists of points of unstable equilibrium.*

Next let B be a point of the line (iii), say, $B(0, 0, \tilde{\omega}_3)$ $(\tilde{\omega}_3<0)$. The constants of integration are $\tilde{p}=\tilde{q}=-\tilde{\omega}_3$. With these constants (8) and (9) reduce to

$$a_3\omega_1^2=(\omega_3-\tilde{\omega}_3)\left[(2b_1-a_1\tilde{\omega}_3)-a_1\omega_3\right], \tag{13}$$

$$a_3\omega_2^2=(\omega_3-\tilde{\omega}_3)\left[(-2b_2-a_2\tilde{\omega}_3)-a_2\omega_3\right]. \tag{14}$$

Let $\frac{b_1}{a_1}<\tilde{\omega}_3<-\frac{b_2}{a_2}$. Then $\tilde{\omega}_3<-\frac{2b_2}{a_2}-\tilde{\omega}_3$. Let I be a closed interval $\left[\tilde{\omega}_3, -\frac{2b_2}{a_2}-\tilde{\omega}_3\right]$. Both factors of (14) are non-negative if $\omega_3\in I$, i.e., ω_2 is real for all values of ω_3 in I. Again since $\frac{b_1}{a_1}<\tilde{\omega}_3$ it follows that $\frac{2b_1}{a_1}-\tilde{\omega}_3<\tilde{\omega}_3$. Hence $\omega_3\in I$ implies that $\frac{2b_1}{a_1}-\tilde{\omega}_3<\omega_3$ or $2b_1-a_1\tilde{\omega}_3>a_1\omega_3$. Therefore for every value of ω_3 in I, ω_1 is real. Since $\tilde{\omega}_3\in I$, it follows by the lemma that B is a point of unstable equilibrium.

Points of stable equilibrium. It follows from (8) and (9) that for $\varepsilon_1>0$ there exists an $\eta>0$ such that

$$|p^2-\tilde{p}^2|<\eta,\qquad |q-\tilde{q}|<\eta, \tag{15}$$

and

$$|\omega_3-\tilde{\omega}_3|<\eta<\varepsilon_1 \tag{16}$$

imply that

$$|\omega_i^2-\tilde{\omega}_i^2|<\varepsilon_1,\qquad i=1,2. \tag{17}$$

In fact, by (6) and (7) for every $\nu>0$ there exists a $\delta>0$ such that

$$|\overline{\omega}_i-\tilde{\omega}_i|<\delta,\qquad i=1,2,3 \tag{18}$$

implies

$$|\overline{p}^2-\tilde{p}^2|<\nu,\qquad |\overline{q}-\tilde{q}|<\nu. \tag{19}$$

If $(\omega_1, \omega_2, \omega_3)$ is a point on an integral curve of (5), then ω_2 defined by (9) must be real, and therefore

$$\frac{-b_2-R_2}{a_2}\leqq\omega_3\leqq\frac{-b_2+R_2}{a_2}, \tag{20}$$

where $R_2^2 = R_2^2(p, q) = b_2^2 + a_2^2 p^2 - 2a_2 b_2 q$; $R_2 \geqq 0$. Also, (20) can be written in the form

(21) $$\left|\omega_3 + \frac{b_2}{a_2}\right| \leqq \frac{R_2}{a_2}.$$

Let E be an arbitrary point on the line (i), say, $E\left(\tilde{\omega}_1, 0, -\frac{b_2}{a_2}\right)$. By substituting the corresponding constants of integration $\tilde{p}^2$ and $\tilde{q}$ in R_2 it is seen that $\tilde{R}_2 = R_2(\tilde{p}, \tilde{q}) = 0$. Hence for every $\eta > 0$ there exists a $\mu > 0$ such that

(22) $$|\bar{p}^2 - \tilde{p}^2| < \mu, \quad |\bar{q} - \tilde{q}| < \mu$$

imply

(23) $$R_2(\bar{p}, \bar{q}) < a_2 \eta.$$

Let $\varepsilon > 0$ be arbitrary. Since for $\tilde{\omega}_1 = 0$ E is a point of the type F, which is discussed below, it can be assumed that $\tilde{\omega}_1 \neq 0$. Let $\varepsilon_1 = \min\left(\frac{1}{2}\tilde{\omega}_1^2, \frac{|\tilde{\omega}_1|\sqrt{\varepsilon}}{3}, \frac{\varepsilon}{9}, \frac{\sqrt{\varepsilon}}{3}\right)$ and choose η in (15) and (16) such that (17) is satisfied. With η so defined choose μ in (22) such that (23) is satisfied. In (19) let $\nu = \min(\mu, \eta)$, and choose $\delta > 0$ in (18) such that $\delta < |\tilde{\omega}_1|$ and that (19) is satisfied. Now let $Q(\bar{\omega}_1, \bar{\omega}_2, \bar{\omega}_3)$ be an arbitrary but fixed point of $U(E, \delta)$. Then (18) is satisfied, and hence (19) holds with $\bar{p}^2$ and $\bar{q}$ being the constants of integration for the motion starting at Q and tracing a certain curve Γ. Since (19) holds, it follows by definition of ν that (22) is satisfied; hence (23) holds. For all points of Γ, p^2 and q have the same values, namely $\bar{p}^2$ and $\bar{q}$. Therefore (23) and (21) imply (16); (15) is implied by (19). Thus, (17) is satisfied for every point of Γ. Since $\tilde{\omega}_2 = 0$, (17) implies that $\omega_2^2 < \frac{\varepsilon}{9}$, or $|\omega_2| = |\omega_2 - \tilde{\omega}_2| < \frac{\sqrt{\varepsilon}}{3}$. Also, by (17), $|\omega_1^2 - \tilde{\omega}_1^2| < \frac{1}{2}\tilde{\omega}_1^2$, i.e., $-\frac{1}{2}\tilde{\omega}_1^2 < \omega_1^2 - \tilde{\omega}_1^2$, or $\frac{1}{2}\tilde{\omega}_1^2 < \omega_1^2$, i.e., $\omega_1^2 \neq 0$ for all points of Γ. Since $|\bar{\omega}_1 - \tilde{\omega}_1| < \delta < |\tilde{\omega}_1|$, $\tilde{\omega}_1 > 0$ implies that $-\tilde{\omega}_1 < \bar{\omega}_1 - \tilde{\omega}_1$ or $\bar{\omega}_1 > 0$; similarly $\tilde{\omega}_1 < 0$ implies that $\bar{\omega}_1 - \tilde{\omega}_1 < -\tilde{\omega}_1$ or $\bar{\omega}_1 < 0$, i.e., $\bar{\omega}_1$ has the same sign as $\tilde{\omega}_1$. Since ω_1 starts from $\bar{\omega}_1$ and varies continuously on Γ, and since $\omega_1^2 \neq 0$ for every point of Γ, it follows that ω_1 has the same sign as $\tilde{\omega}_1$ for every point of Γ. Therefore $|\omega_1 + \tilde{\omega}_1| > |\tilde{\omega}_1|$ for every point of Γ. Hence by (17) $|\omega_1 - \tilde{\omega}_1| < \frac{\varepsilon_1}{|\omega_1 + \tilde{\omega}_1|} < \frac{\varepsilon_1}{|\tilde{\omega}_1|}$. Therefore by definition of ε_1, $|\omega_1 - \tilde{\omega}_1| < \frac{\sqrt{\varepsilon}}{3}$ on Γ. This completes the proof that $\Gamma < U(E, \varepsilon)$. *Hence the points on the line* (i) *with* $\omega_1 \neq 0$ *are the points of stable equilibrium.*

Let F be a point on the line (iii), say, $F(0, 0, \tilde{\omega}_3)$. With the corresponding new interpretation of $\tilde{\omega}_1, \tilde{\omega}_2, \tilde{\omega}_3, \tilde{p}^2, \tilde{q}$ and arbitrary $\varepsilon > 0$, let $\varepsilon_1 = \min\left(\frac{\varepsilon}{9}, \frac{\sqrt{\varepsilon}}{3}\right)$ and choose η in (15) and (16) such that (17) is satisfied. Now $\tilde{p}^2 = \tilde{\omega}_3^2$, $\tilde{q} = -\omega_3$, and $\tilde{R}_2^2 = (b_2 + a_2\tilde{\omega}_3)^2$. Therefore there exists $\mu > 0$ such that if (22) is satisfied, then

(23′) $$|R_2 - |b_2 + a_2\tilde{\omega}_3|| < a_2 \eta.$$

Let $\nu = \min(\mu, \eta)$ and choose $\delta > 0$ in (18) such that (19) is satisfied. Suppose now that the motion starts at some point $(\bar{\omega}_1, \bar{\omega}_2, \bar{\omega}_3) \in U(F, \delta)$. It follows from (7) (since $a_1 < 0$) that for every $t \geqq 0$ we have $\omega_3 \geqq -\bar{q}$; hence $\omega_3 - \tilde{\omega}_3 > -\bar{q} + \tilde{q}$.

By definition of δ the last inequality can be written as

(24) $$\omega_3 - \tilde{\omega}_3 > -\eta.$$

Suppose now that

(25) $$-\frac{b_2}{a_2} \leqq \tilde{\omega}_3,$$

i.e., $b_2 + a_2\tilde{\omega}_3 \geqq 0$. Then it follows from (23′) that $\frac{-b_2+R_2}{a_2} < \tilde{\omega}_3 + \eta$. Hence, by (20), $\omega_3 - \tilde{\omega}_3 < \eta$ for all t. This, together with (24), shows that (16) is satisfied. Since (15) is implied by (19), (17) follows. Since now $\tilde{\omega}_1 = \tilde{\omega}_2 = 0$, (17) implies that $\omega_i^2 < \frac{\varepsilon}{9}$, or $|\omega_i| = |\omega_i - \tilde{\omega}_i| < \frac{\sqrt{\varepsilon}}{3}$. Thus, F is a point of stable equilibrium, *i.e., all points of* (iii) *for which* $\omega_3 \geqq \frac{-b_2}{a_2}$ *are points of stable equilibrium.*

Next consider the case $\tilde{\omega}_3 < \frac{b_1}{a_1} < 0$. Let $R_1^2(p, q) = b_1^2 + a_1^2 p^2 + 2a_1 b_1 q$; then $\tilde{R}_1^2 = R_1^2(\tilde{p}, \tilde{q}) = (b_1 - a_1\tilde{\omega}_3)^2 > 0$, or $\tilde{R}_1 = a_1\tilde{\omega}_3 - b_1 > 0$. There exists $\mu > 0$, such that for every $\bar{p}$ and $\bar{q}$, satisfying (22), $\bar{R}_1^2 > 0$ and

(26) $$|\bar{R}_1 - \tilde{R}_1| < |a_1|\,\eta.$$

With this new definition of μ, and with η defined as above, again let $\nu = \min(\mu, \eta)$. Now choose $\delta > 0$ in (18) such that $\delta < \frac{b_1}{a_1} - \tilde{\omega}_3$ and such that (19) is satisfied. Then (22) and (24) hold; (22) implies that $\bar{R}_1^2 = R_1^2(\bar{p}, \bar{q}) > 0$. [$\bar{p}$ and $\bar{q}$ are as before constants of integration corresponding to $Q(\tilde{\omega}_1, \tilde{\omega}_2, \tilde{\omega}_3) \in U(F, \delta)$.] Therefore it follows from (8) that for every point of the integral curve Γ passing through Q either $\omega_3 \geqq \frac{b_1 - \bar{R}_1}{a_1}$ or $\omega_3 \leqq \frac{b_1 + \bar{R}_1}{a_1}$. But certainly $\bar{\omega}_3 < \tilde{\omega}_3 + \delta < \tilde{\omega}_3 + \frac{b_1}{a_1} - \tilde{\omega}_3$. Since $-\frac{\bar{R}_1}{a_1} > 0$, this implies that $\bar{\omega}_3 \leqq \frac{b_1 - \bar{R}_1}{a_1}$. Since $\bar{\omega}_3$ corresponds to a point of Γ, this implies that $\bar{\omega}_3 \leqq \frac{b_1 + \bar{R}_1}{a_1}$. Since ω_3 varies on Γ continuously, it follows that

(27) $$\omega_3 \leqq \frac{b_1 + \bar{R}_1}{a_1}$$

for every point on Γ. Replacing (26) by $\tilde{R}_1 - \bar{R}_1 < |a_1|\,\eta$ or $a_1\tilde{\omega}_3 - b_1 - \bar{R}_1 < |a_1|\,\eta$, dividing this inequality by $|a_1|$, and substituting from (27), we see that

(28) $$\omega_3 - \tilde{\omega}_3 < \eta.$$

Now (24) and (28) show that (16) is satisfied for every point of Γ. Since (15) is implied by (22), (16) holds for every point of Γ; thus it follows again in the same way as above that $\Gamma \subset U(F, \varepsilon)$, i.e., *the points of* (iii) *for which* $\omega_3 < \frac{b_1}{a_1}$ *are points of stable equilibrium.* In summary of the results above it can be stated that *the points* $\left(\omega_1, 0, -\frac{b_2}{a_2}\right)$ *and* $(0, 0, \omega_3)$ *with* $\omega_3 \geqq -\frac{b_2}{a_2}$ or $\omega_3 < \frac{b_1}{a_1} - k$ *are uniformly stable for every* $k > 0$. *The points in a neighborhood of* $\left(0, 0, \frac{b_1}{a_1}\right)$ *with* $\omega_3 < \frac{b_1}{a_1}$ *are stable, but not uniformly. The points* $(0, 0, \omega_3)$ *with* $\frac{b_1}{a_1} \leqq \omega_3 < -\frac{b_2}{a_2}$ *and the points* $\left(0, \omega_2, \frac{b_1}{a_1}\right)$ *are unstable.*

Case (b). *Unstable points.* If a point B is on the line (iii) and $\omega_3 < -\frac{b_2}{a_2}$, it follows in the same way as in case (a) above that ω_2 is real provided $\omega_3 \in I$. Since now (13) reduces to $a_3\omega_1^2 = \omega_3 + \tilde{p}$, it is seen that ω_1 is real if $\omega_3 \in I$, thus *the points* $(0, 0, \omega_3)$ *with* $\omega_3 < -\frac{b_2}{a_2}$ *are unstable.*

Stable points. The argument of Case (a) can be repeated to show that *all points of the line* (i) *and those points of* (iii) *for which* $\omega_3 \geqq -\frac{b_2}{a_2}$ *are also stable equilibrium points.*

Case (c). *Unstable points.* Let B have coordinates $(0, 0, \tilde{\omega}_3)$. The solution of (5) is again given by (13) and (14). If $\tilde{\omega}_3 < -\frac{b_2}{a_2}$, it follows, as before, that for every $\omega_3 \in I$, ω_2 is real. Since now $\frac{b_1}{a_1} > 0$, it follows that $\frac{2b_1}{a_1} - \tilde{\omega}_3 > -\frac{2b_2}{a_2} - \tilde{\omega}_3$, i.e., $\omega_3 \in I$ implies that $\frac{2b_1}{a_1} - \tilde{\omega}_3 > \omega_3$ whence it follows that ω_1 as defined by (13) is real. Therefore, by the lemma, B is an unstable point. If $\tilde{\omega}_3 > \frac{b_1}{a_1} > 0$, then $2\frac{b_1}{a_1} - \tilde{\omega}_3 < \tilde{\omega}_3$. Hence if $c = \max\left(0, 2\frac{b_1}{a_1} - \tilde{\omega}_3\right)$, it follows that $c < \tilde{\omega}_3$ and therefore a closed interval $I' = [c, \tilde{\omega}_3]$ exists. If $\omega_3 \in I'$, then the first factors of (13) and (14) are never positive. The second factor of (14) is negative for all values of $\omega_3 \geqq 0$. Hence $\omega_3 \in I'$ implies that ω_2 is real. Now, for $\omega_3 \in I'$ we have $\omega_3 \geqq 2\frac{b_1}{a_1} - \tilde{\omega}_3$; hence the second factor of (13) is never positive, *i.e.*, ω_1 is real provided $\omega_3 \in I'$. Therefore the point B is again unstable, by the lemma.

Stable points. Consider $E\left(\tilde{\omega}_1, 0, -\frac{b_2}{a_2}\right)$, i.e., E is a point on the line (i). Since (20) holds also in this case, the argument about the line (i) in Case (a) can be repeated and the same result obtained. If $G\left(0, \tilde{\omega}_2, \frac{b_1}{a_1}\right)$ is a point of the line (ii) and $\tilde{p}, \tilde{q}$ are the corresponding constants of integration, then $R_1(\tilde{p}, \tilde{q}) = 0$. Furthermore (8) now implies

$$\frac{b_1 - R_1}{a_1} \leqq \omega_3 \leqq \frac{b_1 + R_1}{a_1}. \tag{29}$$

The argument of Case (a) about the line (i) can be repeated using this last inequality to lead to the conclusion that *the line* (ii) *consists of stable points.*

Stable points on the line (iii). Consider a point $F(0, 0, \tilde{\omega}_3)$ with $-\frac{b_2}{a_2} \leqq \tilde{\omega}_3 \leqq \frac{b_1}{a_1}$. Since (20) holds in this case as well as in the Case (a), it follows as in Case (a) that

$$\omega_3 - \tilde{\omega}_3 < \eta. \tag{30}$$

Now $b_1 - a_1\tilde{\omega}_3 \geqq 0$; hence $\widetilde{R}_1 = b_1 - a_1\tilde{\omega}_3$, and therefore it follows from (26) that $\frac{b_1 - R_1}{a_1} - \tilde{\omega}_3 > -\eta$, and therefore by (29) that $\omega_3 - \tilde{\omega}_3 > -\eta$. This, together with (30), establishes (17), and the argument can be completed in the same way as for the point F with $\tilde{\omega}_3 \geqq -\frac{b_2}{a_2}$ in case (a). Thus, *in this case the points* $F(0, 0, \tilde{\omega}_3)$ *with* $-\frac{b_2}{a_2} \leqq \tilde{\omega}_3 \leqq \frac{b_1}{a_1}$ *are stable.*

The transformation $\omega_1=\omega_2'$, $\omega_2=\omega_1'$, $\omega_3=\omega_3'$ reduces Cases (d) and (e) to the Cases (b) and (a), respectively. The complete classification of the singular points is given in the table below.

Table. *Equilibrium Points*

Case	Stable	Unstable
(a)	$-\infty<\omega_1<\infty,\quad \omega_2=0,\quad \omega_3=-\frac{b_2}{a_2}$	$\omega_1=0,\quad -\infty<\omega_2<\infty,\quad \omega_3=\frac{b_1}{a_1}$
	$\omega_1=\omega_2=0,\quad \omega_3\geqq-\frac{b_2}{a_2}$	$\omega_1=\omega_2=0,\quad \frac{b_1}{a_1}\leqq\omega_3<-\frac{b_2}{a_2}$
	$\omega_1=\omega_2=0,\quad \omega_3<\frac{b_1}{a_1}$	
(b)	$-\infty<\omega_1<\infty,\quad \omega_2=0,\quad \omega_3=-\frac{b_2}{a_2}$	$\omega_1=\omega_2=0,\quad \omega_3<-\frac{b_2}{a_2}$
	$\omega_1=\omega_2=0,\quad \omega_3\geqq-\frac{b_2}{a_2}$	
(c)	$-\infty<\omega_1<\infty,\quad \omega_2=0,\quad \omega_3=-\frac{b_2}{a_2}$	$\omega_1=\omega_2=0,\quad \omega_3<-\frac{b_2}{a_2}$
	$\omega_1=0,\quad -\infty<\omega_2<\infty,\quad \omega_3=\frac{b_1}{a_1}$	$\omega_1=\omega_2=0,\quad \omega_3>\frac{b_1}{a_1}$
	$\omega_1=\omega_2=0,\quad -\frac{b_2}{a_2}\leqq\omega_3\leqq\frac{b_1}{a_1}$	
(d)	$\omega_1=0,\quad -\infty<\omega_2<\infty,\quad \omega_3=\frac{b_1}{a_1}$	$\omega_1=\omega_2=0,\quad \omega_3>\frac{b_1}{a_1}$
	$\omega_1=\omega_2=0,\quad \omega_3\leqq\frac{b_1}{a_1}$	
(e)	$\omega_1=0,\quad -\infty<\omega_2<\infty,\quad \omega_3=\frac{b_1}{a_1}$	$-\infty<\omega_1<\infty,\quad \omega_2=0,\quad \omega_3=-\frac{b_2}{a_2}$
	$\omega_1=\omega_2=0,\quad \omega_3>-\frac{b_2}{a_2}$	$\omega_1=\omega_2=0,\quad \frac{b_1}{a_1}<\omega_3\leqq-\frac{b_2}{a_2}$
	$\omega_1=\omega_2=0,\quad \omega_3\leqq\frac{b_1}{a_1}$	

4. Complete Solution of the Equations of Motion

Substitution of (8) and (9) in the third equation of (5) yields an equation in ω_3 which can be reduced to quadratures. The solution is expressible in terms of elliptic functions (including the degenerate cases with modulus $\varkappa=1$ or 0). The exact form of a solution depends on the character of the zeros of the right sides of (8) and (9), *i.e.*, on the sign of R_1^2 and R_2^2. Since R_1^2 and R_2^2 are linear in q, it suffices to examine the signs of R_1^2 and R_2^2 at extremal values of q for any fixed value of p. A simple calculation shows that the extreme values of q are:

$$q_1=\frac{a_2}{2b_2}\left(p^2+\frac{b_2^2}{a_2^2}\right),$$

$$q_2=-\frac{a_1}{2b_1}\left(p^2+\frac{b_1^2}{a_1^2}\right),$$

$$q_3=-p,$$

$$q_4=p.$$

Then

$$q_1 - q_2 = \frac{\beta}{2b_1 b_2}\left(p^2 + \frac{b_1 b_2}{a_1 a_2}\right), \tag{31}$$

$$q_1 - q_3 = \frac{a_2}{2b_2}\left(p + \frac{b_2}{a_2}\right)^2, \tag{32}$$

$$q_1 - q_4 = \frac{a_2}{2b_2}\left(p - \frac{b_2}{a_2}\right)^2, \tag{33}$$

$$q_2 - q_3 = \frac{-a_1}{2b_1}\left(p - \frac{b_1}{a_1}\right)^2, \tag{34}$$

$$q_2 - q_4 = \frac{-a_1}{2b_1}\left(p + \frac{b_1}{a_1}\right)^2, \tag{35}$$

$$R_1^2(q_1) = \frac{a_1 \beta}{b_2}\left(p^2 + \frac{b_1 b_2}{a_1 a_2}\right), \tag{36}$$

$$R_1^2(q_2) = 0, \tag{37}$$

$$R_1^2(q_3) = (b_1 - a_1 p)^2, \tag{38}$$

$$R_1^2(q_4) = (b_1 + a_1 p)^2, \tag{39}$$

$$R_2^2(q_1) = 0, \tag{40}$$

$$R_2^2(q_2) = \frac{a_2 \beta}{b_1}\left(p^2 + \frac{b_1 b_2}{a_1 a_2}\right), \tag{41}$$

$$R_2^2(q_3) = (b_2 + a_2 p)^2, \tag{42}$$

$$R_2^2(q_4) = (b_2 - a_2 p)^2. \tag{43}$$

The zeros of the right-hand side of (8) are r_1 and r_2; those of (9) are r_3 and r_4, where

$$r_1 = \frac{b_1 + \sqrt{R_1^2}}{a_1}, \quad r_2 = \frac{b_1 - \sqrt{R_1^2}}{a_1}, \quad r_3 = -\frac{b_2 + \sqrt{R_2^2}}{a_2}, \quad r_4 = -\frac{b_2 - \sqrt{R_2^2}}{a_2}.$$

Case (a). It follows from (6) and (7) that a solution of (5) is real only if

$$-p \leqq q. \tag{44}$$

But (9) and (40) imply that ω_2 is real if and only if

$$q < q_1. \tag{45}$$

Since $a_1 < 0$, it follows from (8) and (9) that ω_1 and ω_2 are both real for properly chosen values of ω_3 if and only if

$$r_3 \leqq r_1 \tag{46}$$

or

$$r_2 \leqq r_4, \tag{47}$$

provided r_1 and r_2 are real. (47) is satisfied only if

$$q \leqq \frac{b_2}{a_2} \quad \text{and} \quad |q| \leqq p, \tag{48}$$

or

$$q > \frac{b_2}{a_2} \quad \text{and} \quad p > \frac{b_2}{a_2}. \tag{49}$$

The second inequality of (49) follows from the first and (45). If $p \leqq \frac{b_2}{a_2}$, then $p^2 + \frac{b_1 b_2}{a_1 a_2} < 0$ and hence by (31) $q_2 < q_1$; by (37) and (45) $R_1^2(q) \geqq 0$, i.e., r_1 and r_2 are real. By (45) $q \leqq \frac{b_2}{a_2}$. Since in this case $a_2 q - b_2 \leqq 0$, one has $a_2 q - b_2 \leqq \sqrt{R_2^2}$ which is equivalent to $\frac{2b_2}{a_2} \cdot \frac{\beta}{a_1 a_2} + \frac{2\beta \sqrt{R_2^2}}{a_1 a_2^2} \leqq \frac{2q\beta}{a_1 a_2}$. By adding $\frac{b_1^2}{a_1^2} + p^2 - \frac{2b_2 q}{a_2}$ to both sides of the above inequality one gets

$$\frac{\beta^2}{a_1^2 a_2^2} + \frac{b_2^2 + a_2^2 p^2 - 2a_2 b_2 q}{a_2^2} + \frac{2\beta\sqrt{R_2^2}}{a_1 a_2^2} \leqq \frac{b_1^2 + a_1^2 p^2 + 2a_1 b_1 q}{a_1^2};$$

that is,

$$\left(\frac{\beta}{a_1 a_2} + \frac{\sqrt{R_2^2}}{a_2}\right)^2 \leqq \left(\frac{\sqrt{R_1^2}}{a_1}\right)^2, \quad \text{or} \quad \frac{\beta}{a_1 a_2} + \frac{\sqrt{R_2^2}}{a_2} \leqq -\frac{\sqrt{R_1^2}}{a_1}.$$

By subtracting b_1/a_1 from both sides of this inequality one gets $-r_3 \leqq -r_1$. Therefore by (46) and (47) in view of (48) and (49) it follows that $|q| \leqq p$. Hence by (39) and (43) $R_1^2(q) > 0$ and $R_2^2(q) > 0$, i.e., $r_3 \neq r_4$, unless $p = \frac{b_2}{a_2} = q$. When $-p < q < p$, all roots are distinct, and $r_3 < r_2$. Then ω_3 is in the form

$$\omega_3 = \frac{A \operatorname{sn}^2(C_1 t + C_2, \varkappa) + B}{C \operatorname{sn}^2(C_1 t + C_2, \varkappa) + D}, \tag{50}$$

where C_2 is a constant of integration and A, B, C, D, and C_1 are known functions of r_1, r_2, r_3, and r_4. The form of these functions depends on the position of the initial value of ω_3 relative to r_1, r_2, r_3, and r_4. If $|q| = p$, then the point $(0, 0, \tilde{\omega}_3)$ with $|\tilde{\omega}_3| = p$ lies on the curve defined by the corresponding solution of (5). Since $p \leqq \frac{b_2}{a_2}$ or $\tilde{\omega}_3 \geqq -\frac{b_2}{a_2}$, by the preceding section $(0, 0, \tilde{\omega}_3)$ is a stable point, and therefore by the lemma the solution is a degenerate curve consisting of a single point $(0, 0, \tilde{\omega}_3)$.

Next consider the case $\frac{b_2}{a_2} < p \leqq \sqrt{-\frac{b_1 b_2}{a_1 a_2}}$. Then by (31) $q_2 \geqq q_1$, and hence by (37) $R_1^2(q_1) \geqq 0$. R_1^2 is a decreasing function of q. By (44) and (45) $-p \leqq q \leqq q_1$; hence $R_1^2(q_1) \geqq 0$ implies that $R_1^2(q) \geqq 0$, while (40) implies that $R_2^2(q) \geqq 0$. Therefore r_1, r_2, r_3, and r_4 are real and distinct except when $q = q_1, q_3, q_4$. Thus, when $q \neq q_1, q_3, q_4$, ω_3 is of the form (50).

If $q = q_3$, then the point $(0, 0, \tilde{\omega}_3)$ with $\tilde{\omega}_3 = p$ lies on the curve defined by the solution corresponding to the constants p and q_3. Since $\tilde{\omega}_3 = -q_3 = p$, it follows that $\tilde{\omega}_3 > \frac{b_2}{a_2}$, hence $\tilde{\omega}_3 > -\frac{b_2}{a_2}$, and the solution is degenerate.

If $q = q_1$, then the coordinates of a point on the line (i) satisfy the solution and again the solution is degenerate, as follows from the lemma, since the line (i) consists of stable points. If $q = q_4$, then $r_2 = r_3 = -p$, and

$$\dot{\omega}_3 = \pm (\omega_3 + p) \sqrt{(2b_1 + a_1 p - a_1 \omega_3)(-2b_2 + a_2 p - a_2 \omega_3)}. \tag{51}$$

The integration yields

$$\omega_3 = \frac{A \tanh^2(C_1 t + C_2) + B}{C \tanh^2(C_1 t + C_2) + D}, \tag{52}$$

where C_2 is a constant of integration and C_1, A, B, C, and D are known functions of r_1, r_4, and p.

If $\sqrt{-\frac{b_1 b_2}{a_1 a_2}} < p$, then by (31) $q_2 < q_1$. Since, as before, the range of q is between $-p$ and q_1, it follows from (37) that $R_1^2(q)$ changes sign when $q = q_2$. By (40) $R_2^2(q) \geqq 0$. Therefore for $q < q_2$ and q not one of q_3, q_4, all roots r_1, r_2, r_3, and r_4 are real and distinct. Accordingly, this leads to a solution of the form (50). If $q = q_3$, it follows again in the same way as above that the solution is degenerate. The same holds for $q = q_1$. If $q = q_4$, the argument leading to (52) requires no modification, *i.e.*, ω_3 is given by (52). If $q = q_2$, then $r_1 = r_2 = \frac{b_1}{a_1}$. Therefore

$$\dot{\omega}_3 = \pm\left(\omega_3 - \frac{b_1}{a_1}\right)\sqrt{-a_1\left(-a_2\,\omega_3^2 - 2b_2\,\omega_3 + \frac{\beta}{b_1}p^2 + \frac{b_1 b_2}{a_1}\right)}. \tag{53}$$

If $p < -\frac{b_1}{a_1}$, integration of (53) yields

$$\omega_3 = \frac{A\tan^2(C_1 t + C_2) + B}{C\tan^2(C_1 t + C_2) + D}, \tag{54}$$

where A, B, C, D, C_1 and C_2 have meanings similar to those in (52). If $p = -\frac{b_1}{a_1}$, then (53) can be written in the following form:

$$\dot{\omega}_3 = \pm\left(\omega_3 - \frac{b_1}{a_1}\right)^{\frac{3}{2}}\sqrt{a_1 a_2\left(\omega_3 + \frac{2b_2}{a_2} + \frac{b_1}{a_1}\right)}. \tag{55}$$

Integration of (55) yields

$$\omega_3 = \frac{A(C_1 t + C_2)^2 + B}{C(C_1 t + C_2)^2 + D}. \tag{56}$$

If $p > -\frac{b_1}{a_1}$, then the integration of (53) leads to (52). If $q_2 < q < q_1$, then all roots are distinct and $R_1^2 < 0$, i.e., r_1 and r_2 are imaginary. In this case

$$\omega_3 = \frac{A\,\mathrm{cn}(C_1 t + C_2) + B}{C\,\mathrm{cn}(C_1 t + C_2) + D}. \tag{57}$$

Case (b). Here $\dot{\omega}_3^2$ is a third-degree polynomial in ω_3. (9) implies that the solution is real only if all three roots of this polynomial are real. The roots are distinct if q is none of q_1, q_3, q_4, and in this case the solution is of the form (50) (for certain regions of the initial values of ω_3 one of the constants C or D is zero). If $q = q_1$, q_3 or $q = q_4$ and $p \leqq \frac{b_2}{a_2}$, then by using the lemma and the results of preceding section it can be shown that solution is degenerate. If $q = q_4$ and $p > \frac{b_2}{a_2}$, then $\omega_3 = A\tanh^2(C_1 t + C_2)$ which is of the form (51).

Case (c). Since in this case $a_1 > 0$ and $a_2 > 0$, then (8) and (9) show that for a real solution all four roots r_1, r_2, r_3, r_4 must be real. For fixed p now R_1^2 is an increasing function while R_2^2 is a decreasing function of q. Hence it follows from (37) and (40) that $q_2 \leqq q \leqq q_1$ and that $r_1 = r_2$ only if $q = q_2$, $r_3 = r_4$ if $q = q_1$. A simple computation shows that any other combination of equal roots occurs only if $q = q_3$ or q_4. Therefore if $q \neq q_i$, $i = 1, 2, 3, 4$, then ω_3 is given by (50). If $q = q_1$ or q_2, then by using the lemma and the fact that the lines (i) and (ii) consist of stable points it can be shown in the same way as above that the solution is degenerate. The same argument applies to $q=q_3$, provided $p \leqq \frac{b_1}{a_1}$, and to $q=q_4$, provided $p \leqq \frac{b_2}{a_2}$. If, however, $q=q_3$ and $p > \frac{b_1}{a_1}$ or $q=q_4$ and $p < \frac{b_2}{a_2}$, then the solution is in the form (51).

Solutions of Cases (d) and (e) can be obtained from the solutions of Cases (b) and (a), respectively, by the transformation mentioned in the preceding section. When ω_3 is expressed as a function of time, the corresponding expressions for ω_1 and ω_2 can be obtained from (8) and (9), respectively.

References

[1] Rogers, E. E.: A Mathematical Model for Predicting the Damping Time of Mercury Damper. U.S.Naval Ordnance Test Station, 1959.
[2] Haseltine, W. R.: Passive Damping of Wobbling Satellites. Part I. General Stability Theory and First Examples. NAVORD Report 6579, Part I, 1959.
[3] Zaroodny, S., & J. W. Bradley: Nutation Damper — A Simple Two-Body Gyroscopic System. Forthcoming as a BRL report, Aberdeen Proving Ground, Maryland.

Ballistic Research Laboratories
Aberdeen Proving Ground
Aberdeen, Maryland

(Received March 18, 1961)

Optimal Control for Nonlinear Processes

E. B. LEE & L. MARKUS

Communicated by J. SERRIN

Introduction

In an optimum control problem we are given a real ordinary differential equation system

$$\frac{dx^i}{dt} = f^i(t; x^1, x^2, \ldots, x^n; u^1, \ldots, u^m) \qquad i = 1, 2, \ldots, n$$

which is a mathematical model of some physical process. The problem of control is to select the real functions $u^j(t)$, $j=1, 2, \ldots, m$ (control variables) on an interval of time, $t_0 \leqq t \leqq t_1$, such that the solution $x^i(t)$ moves in a prescribed manner on $t_0 \leqq t \leqq t_1$. The quality of this choice is measured in terms of a performance index. For example, it could be required that the $x^i(t)$ move from a prescribed initial point to a prescribed moving target $G^i(t)$ in a minimum interval of time by choosing the $u^j(t)$ from an appropriate class of controls; or it could be required that the $x^i(t)$ move to $G^i(t)$ in a finite interval of time by using $u^j(t)$ in which the energy for control is to be a minimum. Usually the performance is measured by a functional which depends on the control variables $u^j(t)$ and the controlled variables $x^i(t)$.

We shall consider the problem of existence of an optimal control. This problem has been solved for the case of linear differential equations in [*2*], [*3*], [*9*], and for certain nonlinear equations in [*8*]; however, our treatment includes these earlier results. In Section 1, Theorems 1 and 2 state conditions such that, if there exists one allowable control which does the prescribed task, an optimal control will exist. The results of Section 2 are concerned with establishing the existence of an allowable control which accomplishes the task for various forms of the differential equation system.

1. Existence of an Optimal Control

Consider the differential system

$$1) \qquad \frac{dx^i}{dt} = f^i(t, x^1, \ldots, x^n, u^1, \ldots, u^m); \qquad i = 1, 2, \ldots, n$$

where[*] $f^i(t, x^1, \ldots, x^n, u^1, \ldots, u^m) = f^i(t, x, u)$ together with

$$\frac{\partial f^i}{\partial x^k}(t, x, u); \qquad i, k = 1, 2, \ldots, n$$

---

[*] In vector notation $x = (x^1, \ldots, x^n)$ and $|x| = \sum_{i=1}^{n} |x^i|$.

are real continuous functions in $R^1 \times R^n \times \Omega$, where R^n is the real n-dimensional number space and Ω is a non-empty compact subset of R^m.

For each choice of the function

$$u(t) = \bigl(u^1(t), \ldots, u^m(t)\bigr) \quad \text{on} \quad -\infty < t_0 \leqq t \leqq t_1 < \infty,$$

as a measurable vector-valued function with a graph in $\Omega \subset R^m$, the differential system

$$1)_{u(t)} \qquad \dot{x}^i = f^i\bigl(t, x, u(t)\bigr) \qquad i = 1, 2, \ldots, n$$

has a unique absolutely continuous solution, or response, $x(t)$ on a subinterval of $t_0 \leqq t \leqq t_1$, with a prescribed initial point $x_0 = x(t_0)$. This is the conclusion of the CARATHÉODORY existence theorem for differential systems [*4*]. Note that $u(t)$ is measurable if and only if every component $u^1(t), \ldots, u^m(t)$ is a real-valued (Lebesgue) measurable function. The response $x(t)$ is continuous, and it has a derivative, except on a set of measure zero, such that the differential system $1)_{u(t)}$ is satisfied almost everywhere. If the finite interval $t_0 \leqq t \leqq t_1$ is degenerate so $t_0 = t_1$, then the response is just the single point $x(t_0) = x_0$.

Definition. A control (or steering function) for system 1) with prescribed non-empty compact set $\Omega \subset R^m$ and prescribed initial point $x_0 \in R^n$ is a measurable vector-valued function $u(t)$, on a finite interval $t_0 \leqq t \leqq t_1$, with $u(t) \subset \Omega$, such that the response $x(t)$ with $x(t_0) = x_0$ is also defined in R^n on $t_0 \leqq t \leqq t_1$.

We shall be interested in those controls such that the response $x(t)$ travels from the prescribed initial point $x_0 = x(t_0)$ to a given moving target $G(t)$. For each t on a given finite interval $\tau_0 \leqq t \leqq \tau_1$, we specify a non-empty compact target set $G(t) \subset R^n$. Moreover, $G(t)$ varies continuously with t. Here we use the Hausdorff metric distance between two non-empty compact subsets X and Y of R^n which is the smallest real number $d = d(X, Y)$ such that X lies in the d-neighborhood of Y and Y lies in the d-neighborhood of X, *cf.* [*1*]. If $G(t)$ is a point for each t, then the target is a continuous curve. If $G(t)$ is a constant compact set, we have the regulator problem where the target is fixed, *cf.* [*10*].

Let us give $f^0(t, x^1, \ldots, x^n, u^1, \ldots, u^m)$ as a real continuous function on $R^1 \times R^n \times \Omega$ and define the cost functional of a control $u(t)$ on $t_0 \leqq t \leqq t_1$, with response $x(t)$, by

$$C(u) = \int_{t_0}^{t_1} f^0\bigl(t, x(t), u(t)\bigr)\, dt.$$

If $f^0(t, x, u) \equiv 1$, then $C(u) = t_1 - t_0$, and the cost of a control is just the time duration over which it acts.

Definition. Given the control problem

a) $\dot{x}^i = f^i(t, x, u), \qquad i = 1, 2, \ldots, n,$
b) $\Omega \subset R^m$,
c) $x_0 \in R^n$,
d) $G(t) \subset R^n$ on $\tau_0 \leqq t \leqq \tau_1$,
e) $C(u)$, the cost functional,

as above. Define $\Delta=\Delta\left(f^1(t, x, u), \ldots, f^n(t, x, u), \Omega, x_0, G(t)\right)$ as the set of all controls $u(t)$, on various subintervals $t_0 \leqq t \leqq t_1$ with $\tau_0 \leqq t_0 \leqq t_1 \leqq \tau_1$, such that

$$x(t_0)=x_0 \quad \text{and} \quad x(t_1) \in G(t_1).$$

A control $u^*(t)$ in Δ is called optimal in case

$$C(u^*) \leqq C(u)$$

for every control $u(t)$ in Δ.

We now prove the basic existence theorem for optimal controls. The examples following the theorem show that an optimal control need not exist if the hypotheses of Theorem 1 are not upheld.

Theorem 1. *Given the control problem*

a) $\dot{x}^i=f^i(t, x^1, \ldots, x^n, u^1, \ldots, u^m)=g^i(t, x)+h_j^i(t, x) u^j$ *for* $i=1, \ldots, n$ *and* $j=1, \ldots, m$ *with* $g^i(t, x)$, $h_j^i(t, x)$, *and* $\frac{\partial g^i}{\partial x^k}(t, x)$, $\frac{\partial h_j^i}{\partial x^k}(t, x)$, $k=1, \ldots, n$, *continuous on* $R^1 \times R^n$,

b) a non-empty, convex, compact restraint set $\Omega < R^m$,

c) the initial point $x_0 \in R^n$,

d) the continuously moving non-empty compact target set $G(t)$ *on the finite interval* $\tau_0 \leqq t \leqq \tau_1$,

e) the cost functional

$$C(u)=\int_{t_0}^{t_1} f^0\left(t, x(t), u(t)\right) dt,$$

where $f^0(t, x, u)=g^0(t, x)+h_j^0(t, x) u^j$, *and* $g^0(t, x)$, *and* $h_j^0(t, x)$ *are continuous on* $R^1 \times R^n$.

Assume the set Δ *of controls with responses traveling from* x_0 *to* G, *as defined above, is such that:*

A) Δ *is non-empty,*

B) there exists a real bound $B<\infty$ *for all responses* $x(t)$ *corresponding to* Δ, *that is,* $|x(t)|<B$ *uniformly for all responses.*

Then there exists an optimal control in Δ.

Proof. Since Δ is non-empty and the corresponding responses are uniformly bounded, $\inf C(u)=m>-\infty$ for all $u \in \Delta$. Either Δ is a finite set, in which case the theorem is trivially true, or we can select a sequence of controls

$$u^{(k)}(t), \quad \text{on} \quad t_0^{(k)} \leqq t \leqq t_1^{(k)} \quad \text{from } \Delta,$$

with $C(u^{(k)})$ decreasing monotonically to m. Select a subsequence (without changing the notation) such that

$$t_0^{(k)} \rightarrow t_0^* \quad \text{and} \quad t_1^{(k)} \rightarrow t_1^*, \quad \text{monotonically.}$$

Let us take the case where

$$t_0^{(k)} \leqq t_0^* \leqq t_1^* \leqq t_1^{(k)} \quad \text{for} \quad k=1, 2, 3, \ldots$$

and consider the other cases later. Then each $u^{(k)}(t)$ is bounded and measurable on the interval $t_0^* \leqq t \leqq t_1^*$ and thus belongs to the Hilbert space $L_2[t_0^*, t_1^*]$. We assume $t_0^*<t_1^*$; for if $t_0^*=t_1^*$, then $m=0$ and $x_0 \in G(t_1^*)$ so every choice of control is optimal.

A closed ball in Hilbert space is weakly compact [5], and thus we can select a further subsequence $u^{(k)}(t)$ with

$$u^{(k)}(t) \to u^*(t) \quad \text{weakly in} \quad L_2[t_0^*, t_1^*].$$

We next show that

$$u^*(t) \quad \text{on} \quad t_0^* \leqq t \leqq t_1^*$$

belongs to Δ.

Now a compact convex set $\Omega \subset R^m$ is precisely the intersection of a finite or a countable number of closed half-spaces, *cf.* [1]. Let

$$a_i\, y^i + b \geqq 0$$

be one such closed half-space in R^m. Let M be the subset of $[t_0^*, t_1^*]$ for which $u^*(t)$ lies in $a_i\, y^i + b < 0$.

If M has positive measure,

$$\int_{t_0^*}^{t_1^*} [a_i\, u^{*\,i}(t) + b]\, \varphi_m(t)\, dt < 0$$

where φ_m is the characteristic function of M. But

$$\lim_{k\to\infty} \int_{t_0^*}^{t_1^*} [a_i\, u^{(k)\,i}(t) + b]\, \varphi_m(t)\, dt = \int_{t_0^*}^{t_1^*} [a_i\, u^{*\,i}(t) + b]\, \varphi_m(t)\, dt.$$

This is impossible since the left members are each non-negative. Thus M has measure zero. Since there are only a countable number of closed half-spaces considered, $u^*(t)$ lies in Ω except on a set of measure zero. Redefine $u^*(t)$ on the exceptional null set of $t_0^* \leqq t \leqq t_1^*$ so that $u^*(t)$ lies everywhere in Ω.

Next consider the response for $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$. The response for

$$u^{(k)}(t) \quad \text{on} \quad t_0^{(k)} \leqq t \leqq t_1^{(k)}$$

is $x^{(k)}(t)$. Then (using vector notation)

$$x^{(k)}(t) = x_0 + \int_{t_0^{(k)}}^{t} \left[g\bigl(s, x^{(k)}(s)\bigr) + h\bigl(s, x^{(k)}(s)\bigr)\, u^{(k)}(s)\right] ds.$$

Select a further subsequence of controls so that

$$\begin{aligned}
&\lim_{k\to\infty} x^{(k)}(t) = x^*(t) && \text{weakly on} \quad [t_0^*, t_1^*],\\
&\lim_{k\to\infty} g\bigl(t, x^{(k)}(t)\bigr) = g^*(t) && \text{weakly on} \quad [t_0^*, t_1^*],\\
&\lim_{k\to\infty} h\bigl(t, x^{(k)}(t)\bigr) = h^*(t) && \text{weakly on} \quad [t_0^*, t_1^*],\\
&\lim_{k\to\infty} h\bigl(t, x^{(k)}(t)\bigr)\, u^{(k)}(t) = \varphi^*(t) && \text{weakly on} \quad [t_0^*, t_1^*].
\end{aligned}$$

Then, for each fixed t on $t_0^* \leqq t \leqq t_1^*$,

$$\begin{aligned}
\lim_{k\to\infty} x^{(k)}(t) = x_0 + \lim_{k\to\infty} \int_{t_0^{(k)}}^{t_0^*} \left[g\bigl(s, x^{(k)}(s)\bigr) + h\bigl(s, x^{(k)}(s)\bigr)\, u^{(k)}(s)\right] ds +\\
+ \lim_{k\to\infty} \int_{t_0^*}^{t} \left[g\bigl(s, x^{(k)}(s)\bigr) + h\bigl(s, x^{(k)}(s)\bigr)\, u^{(k)}(s)\right] ds.
\end{aligned}$$

Thus

$$\lim_{k\to\infty} x^{(k)}(t) = x_0 + \int_{t_0^*}^{t} [g^*(s) + \varphi^*(s)]\, ds.$$

Therefore

$$\lim_{k\to\infty} x^{(k)}(t) = \hat{x}(t)$$

exists, for each fixed t. Also $\hat{x}(t)$ is absolutely continuous on $[t_0^*, t_1^*]$ and $\hat{x}(t_0^*) = x_0$.

Using the Lebesgue convergence theorem [*6*], we find

$$\int_{t_0^*}^{t} \hat{x}(s)\, ds = \int_{t_0^*}^{t} x^*(s)\, ds,$$

so $\hat{x}(t) = x^*(t)$ for almost all t on $[t_0^*, t_1^*]$. Changing the definition of $x^*(t)$ to be precisely $\hat{x}(t)$, we now show that $x^*(t)$ is the response to the control $u^*(t)$.

Now

$$x^*(t) = x_0 + \lim_{k\to\infty} \int_{t_0^*}^{t} \left[g\left(s, x^{(k)}(s)\right) + h\left(s, x^{(k)}(s)\right) u^{(k)}(s)\right] ds,$$

so

$$\begin{aligned} x^*(t) = x_0 + \int_{t_0^*}^{t} g\left(s, x^*(s)\right) ds\, + \\ + \lim_{k\to\infty} \int_{t_0^*}^{t} \Big[h\left(s, x^{(k)}(s)\right) u^{(k)}(s) - h\left(s, x^*(s)\right) u^{(k)}(s)\, + \\ + h\left(s, x^*(s)\right) u^{(k)}(s) - h\left(s, x^*(s)\right) u^*(s) + h\left(s, x^*(s)\right) u^*(s)\Big] ds. \end{aligned}$$

Since $u^{(k)}(s) \in \Omega$, which is compact, and since $h\left(s, x^{(k)}(s)\right) \to h\left(s, x^*(s)\right)$ almost uniformly by EGOROFF's theorem [*6*], we find

$$x^*(t) = x_0 + \int_{t_0^*}^{t} \left[g\left(s, x^*(s)\right) + h\left(s, x^*(s)\right) u^*(s)\right] ds.$$

Therefore $x^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ is the response to the control $u^*(t)$.

Now

$$x^{(k)}(t_1^{(k)}) \in G(t_1^{(k)}) \quad \text{for each} \quad k = 1, 2, 3, \ldots.$$

So

$$x^*(t_1^*) = \lim_{k\to\infty} \left[x^{(k)}(t_1^*) - x^{(k)}(t_1^{(k)}) + x^{(k)}(t_1^{(k)})\right],$$

and

$$x^*(t_1^*) = \lim_{k\to\infty} \left[x^{(k)}(t_1^{(k)})\right].$$

If $x^*(t_1^*)$ were not in $G(t_1^*)$, then there would exist a neighborhood N of the compact set $G(t_1^*)$, so that $x^*(t_1^*)$ is not in the closure of N. But $G(t) \subset N$ for t sufficiently near t_1^*. Thus $x^{(k)}(t_1^{(k)}) \in N$ for large k and yet $x^*(t_1^*)$ is not in $\overline{N}$. This is a contradiction, and therefore $x^*(t_1^*) \in G(t_1^*)$, and the control $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ belongs to Δ.

Now compute the cost of $u^*(t)$. Here

$$C(u^{(k)}) = \int_{t_0^{(k)}}^{t_1^{(k)}} \left[g^0\left(t, x^{(k)}(t)\right) + h_j^0\left(t, x^{(k)}(t)\right) u^{(k)j}(t)\right] dt$$

and

$$\lim_{k\to\infty} C(u^{(k)}) = \int_{t_0^*}^{t_1^*} g^0\big(t, x^*(t)\big)\, dt + \lim_{k\to\infty} \int_{t_0^*}^{t_1^*} h_j^0\big(t, x^{(k)}(t)\big)\, u^{(k)j}(t)\, dt.$$

Just as above we compute

$$\lim_{k\to\infty} C(u^{(k)}) = C(u^*) = m.$$

Therefore $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ is an optimal control.

We return to the assumption

$$t_0^{(k)} \to t_0^*, \qquad t_1^{(k)} \to t_1^* \quad \text{monotonically.}$$

Suppose we do not have

$$t_0^{(k)} \leqq t_0^* \leqq t_1^* \leqq t_1^{(k)}$$

but instead, for example,

$$t_0^{(k)} \leqq t_0^* \quad \text{and} \quad t_1^{(k)} \leqq t_1^*,$$

and the other cases can be treated similarly. Extend each control $u^{(k)}(t)$ (at least for all large k) to the interval $t_0^* \leqq t \leqq t_1^*$ by defining $u^{(k)}(t) = u_0$, a constant vector in Ω, for $t_1^{(k)} \leqq t \leqq t_1^*$. Again define the weak limit in Ω,

$$u^*(t) = \lim_{k\to\infty} u^{(k)}(t) \quad \text{on} \quad t_0^* \leqq t \leqq t_1^*.$$

We must show that each response

$$x^{(k)}(t) \quad \text{on} \quad t_0^{(k)} \leqq t \leqq t_1^{(k)}$$

can be extended to the interval $t_0^{(k)} \leqq t \leqq t_1^*$ using the extended controls. Then we shall show that $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ is in Δ and is an optimal control.

It is easily seen that all the compact sets $G(t)$ on $\tau_0 \leqq t \leqq \tau_1$ lie within one sphere, $S(0, \varrho)$ of radius ϱ, centered at the origin. Thus, for $(t_1^* - t_1^{(k)})$ sufficiently small, each $x^{(k)}(t_1^{(k)})$ lies in $S(0, \varrho)$ and has an extended response on $t_0^{(k)} \leqq t \leqq t_1^*$ which lies in $S(0, 2\varrho)$. Also

$$\lim_{k\to\infty} \big| x^{(k)}(t_1^{(k)}) - x^{(k)}(t_1^*) \big| = 0.$$

Just as above we find that

$$\lim_{k\to\infty} x^{(k)}(t) = x^*(t)$$

at each t on $t_0^* \leqq t \leqq t_1^*$, and moreover $x^*(t)$ is the absolutely continuous response to the control $u^*(t)$, and $x^*(t_0^*) = x_0$.

Now

$$x^{(k)}(t_1^{(k)}) \in G(t_1^{(k)})$$

and

$$\lim_{k\to\infty} \big| x^*(t_1^*) - x^*(t_1^{(k)}) \big| = 0.$$

Thus

$$x^*(t_1^*) = \lim_{k\to\infty} x^*(t_1^{(k)}) = \lim_{k\to\infty} \big[x^{(k)}(t_1^{(k)}) - x^{(k)}(t_1^{(k)}) + x^*(t_1^{(k)}) \big]$$

or

$$x^*(t_1^*) = \lim_{k\to\infty} x^{(k)}(t_1^{(k)}).$$

Thus

$$x^*(t_1^*) \in G(t_1^*),$$

as required. Therefore, $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ lies in Δ.

Finally,

$$C(u^{(k)}) = \int_{t_0^{(k)}}^{t_1^{(k)}} \left[g^0\left(t, x^{(k)}(t)\right) + h_j^0\left(t, x^{(k)}(t)\right) u^{(k)j}(t)\right] dt$$

approaches the limit m as $k \to \infty$. As above we compute

$$\lim_{k \to \infty} C(u^{(k)}) = C(u^*) = m,$$

and thus $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ is an optimal control. Q.E.D.

Remarks. Consider $\Delta(t_0)$, the subset of Δ for which the control $u(t)$ and the response $x(t)$ initiate at a fixed t_0. If $\Delta(t_0)$ is non-empty, and if the responses $x(t)$ for control in $\Delta(t_0)$ are uniformly bounded, then there exists a control in $\Delta(t_0)$ which is optimal relative to $\Delta(t_0)$. The same applies to the set $\Delta(t_0, t_1) \subset \Delta(t_0)$ where the control time interval $t_0 \leqq t \leqq t_1$ is fixed. If the differential system and the cost functional integrand are time-independent, each control in Δ has the same response as some control in $\Delta(\tau_0)$, after a time translation.

If $C(u) > m > -\infty$ for all controls u in Δ, or in $\Delta(t_0)$, then one requires only the uniform bound $|x(t)| < B$ for responses $x(t)$ corresponding to controls $u(t)$ with $C(u)$ near m.

Also $f^i(t, x, u)$ need only be defined and satisfy the hypotheses of the theorem, for $\tau_0 \leqq t \leqq \tau_1$, $x \in \mathcal{O} \subset R^n$, and $u \subset \Omega \subset R^m$, where $\mathcal{O}$ is an open set in R^n which contains the initial point x_0, the moving target $G(t)$, and all the responses of Δ, or of $\Delta(t_0)$, in a compact subset.

The hypothesis A) of the theorem concerns the domain of controllability for the problem, as will be discussed in a later section. The hypothesis B) is satisfied if

$$|f^i(t, x, u)| < \alpha, \qquad i = 1, 2, \ldots, n$$

or if

$$\left|\frac{\partial f^i}{\partial x^k}(t, x, u)\right| < \alpha, \qquad i, k = 1, 2, \ldots, n,$$

for some real α, in $[\tau_0, \tau_1] \times R^n \times \Omega$. Thus B) is always satisfied if $g^i(t, x)$ and $h_j^i(t, x)$ are linear in x.

The following examples illustrate situations where the optimal control fails to exist or is not unique.

Example 1.

$$\dot{x} = \sin 2\pi u, \quad \dot{y} = \cos 2\pi u, \quad \dot{z} = -1 \quad \text{in } R^3.$$

The initial point is $(0, 0, 1)$, and the target is the fixed point $(0, 0, 0)$ on the time interval $0 \leqq t \leqq t_1 \leqq 2$. The restraint set Ω is $-1 \leqq u \leqq 1$. The cost functional is $C(u) = \int_0^{t_1} (x^2 + y^2)\, dt$, and we consider the set of controls $\Delta(0)$.

Each control $u(t)$ in $\Delta(0)$ is defined on $0 \leqq t \leqq 1$. Consider the controls $u^{(k)}(t)$ such that

$$\sin 2\pi u^{(k)}(t) = \sin 2\pi k t,$$
$$\cos 2\pi u^{(k)}(t) = \cos 2\pi k t, \quad \text{for} \quad k = 1, 2, 3, \ldots.$$

Such piecewise continuous controls are easily constructed. The corresponding responses are

$$x^{(k)}(t) = \frac{1 - \cos 2\pi k t}{2\pi k}, \qquad y^{(k)}(t) = \frac{\sin 2\pi k t}{2\pi k}, \qquad z^{(k)}(t) = 1 - t.$$

Thus $x^{(k)}(1) = 0$, $y^{(k)}(1) = 0$, $z^{(k)}(1) = 0$. The cost for each $u^{(k)}(t)$ is computed to be

$$C(u^{(k)}) = \int_0^1 \frac{1 - \cos 2\pi k t}{2\pi^2 k^2}\, dt = \frac{1}{2\pi^2 k^2}.$$

Thus

$$\lim_{k \to \infty} C(u^{(k)}) = 0,$$

and $m = 0$ is the infimum for all $C(u)$ with u in $\Delta(0)$. Yet there is no optimal $u^*(t)$ on $0 \leqq t \leqq 1$ for which the cost is

$$C(u^*) = \int_0^1 (x^{*2} + y^{*2})\, dt = 0.$$

For such an optimal control $u^*(t)$ the response needed is $x^*(t) = 0$, $y^*(t) = 0$. This implies

$$\sin 2\pi u^*(t) = 0 \quad \text{and} \quad \cos 2\pi u^*(t) = 0$$

for almost all t. But this is impossible, and hence there does not exist an optimal control for this control problem. We note that the coefficient functions of the differential equation are not linear in u, but hypotheses A) and B) of the theorem are satisfied.

Example 2.

$$\dot{x} = u_1, \qquad \dot{y} = u_2, \qquad \dot{z} = -1 \quad \text{in } R^3.$$

The initial point is $(0, 0, 1)$, and the target is the fixed point $(0, 0, 0)$ on the time interval $0 \leqq t \leqq t_1 \leqq 2$. Take Ω as the compact but non-convex circle $u_1^2 + u_2^2 = 1$ in R^2. Again take $C(u) = \int_0^{t_1} (x^2 + y^2)\, dt$, and consider the controls $\Delta(0)$. Using the controls $u_1^{(k)} = \cos 2\pi k t$, $u_2^{(k)} = \sin 2\pi k t$, $k = 1, 2, \ldots$, we find $\inf C(u) = 0$ for $u \in \Delta(0)$. Yet there is no optimal control in $\Delta(0)$ which yields a cost of zero.

Example 3.

$$\dot{x} = 1, \qquad \dot{y} = -x e^y u \quad \text{in } R^2.$$

The initial point is $(-1, 0)$, and the target is the fixed point $(1, 0)$ on the time interval $0 \leqq t \leqq t_1 \leqq 2$. The restraint set Ω is $0 \leqq u \leqq 2$. The cost functional is $C(u) = \int_0^{t_1} (2 - y)\, dt = \int_{-1}^{1} (2 - y)\, dx$. Each control $u(t)$ in $\Delta(0)$ is defined on $0 \leqq t \leqq 2$ and yields a response $x(t) = t - 1$, $y(t)$.

Every response $y(t)$ satisfies the inequalities

$$0 \leqq y(x) \leqq -\ln x^2, \quad \text{for} \quad x \neq 0.$$

Each continuous response joining $(-1, 0)$ to $(1, 0)$ must lie below the curve $y = -\ln x^2$ on some interval.

Thus $C(u) > \int\limits_{-1}^{1} (2 + \ln x^2)\, dx = 0$ for each $u(t) \in \Delta(0)$. But for $u(t) = u_\varepsilon = 2 - \varepsilon$ on $0 \leqq t \leqq 2$, we compute the response $x = t - 1$, and

$$y(x) = -\ln\left[\frac{(x^2 - 1)}{2}(2 - \varepsilon) + 1\right].$$

The cost for such a control in $\Delta(0)$ is

$$C(u_\varepsilon) = \int\limits_{-1}^{1} \left\{2 + \ln\left[\frac{(x^2 - 1)}{2}(2 - \varepsilon) + 1\right]\right\} dx.$$

Thus

$$\lim_{\varepsilon \to 0} C(u_\varepsilon) = 0.$$

Hence $\inf C(u) = m = 0$ for $u(t) \in \Delta(0)$. Therefore there does not exist an optimal control in $\Delta(0)$. Here we note that hypothesis B) of the theorem does not hold.

Example 4.

$$\dot{x} = 1, \quad \dot{y} = -x u \quad \text{in } R^2.$$

The initial point is $(-1, 0)$, and the target is the fixed point $(1, 0)$ on the time interval $0 \leqq t \leqq t_1 \leqq 2$. The restraint set Ω is $-1 \leqq u \leqq 1$. The cost functional is $C(u) = \int\limits_0^{t_1} \frac{1}{1 + y^2}\, dt = \int\limits_{-1}^{1} \frac{dx}{1 + y^2}$. Each control $u(t)$ in $\Delta(0)$ is defined on $0 \leqq t \leqq 2$ and yields a response $x(t) = t - 1$, $y(t)$.

Every response $y(t)$ satisfies the inequalities

$$\frac{-(1 - x^2)}{2} \leqq y(x) \leqq \frac{1 - x^2}{2} \quad \text{on} \quad -1 \leqq x \leqq 1.$$

The two controls $u_+(t) = +1$ and $u_-(t) = -1$ are each optimal and achieve the minimal cost. Here an optimal control in $\Delta(0)$ exists, since the hypotheses of Theorem 1 are satisfied, but it is not unique.

We close this section with an existence theorem for Lipschitz continuous controls, which is valid even if the control u enters the coefficients $f(t, x, u)$ in a nonlinear manner.

Theorem 2. *Given the control problem*

a) $\dot{x}^i = f^i(t, x^1, \ldots, x^n, u^1, \ldots, u^m)$; $i = 1, 2, \ldots, n$ *where* $f^i(t, x, u)$ *and* $\frac{\partial f^i}{\partial x^k}(t, x, u)$; $i, k = 1, 2, \ldots, n$ *are continuous in* $R^1 \times R^n \times \Omega$,

b) a non-empty compact restraint set $\Omega \subset R^m$,

c) the initial point $x_0 \in R^n$,

d) the continuously moving non-empty compact target set $G(t) \subset R^n$ *for each* t *on the finite interval* $\tau_0 \leqq t \leqq \tau_1$, *e) the cost functional*

$$C(u) = \int_{t_0}^{t_1} f^0\big(t, x(t), u(t)\big)\, dt,$$

where $f^0(t, x, u)$ *is continuous in* $R^1 \times R^n \times \Omega$.

For a given positive constant A *consider the class* $\Delta(\mathrm{Lip}\, A) \subset \Delta$ *of controls* $u(t)$, *each continuous and satisfying a Lipschitz condition*

$$|u(t) - u(t')| \leq A\,|t - t'|$$

for all pairs t, t' *on some interval* $\tau_0 \leqq t_0 \leqq t$, $t' \leqq t_1 \leqq \tau_1$. *Assume*

A) $\Delta(\mathrm{Lip}\, A)$ *is non-empty*

B) there exists a bound $B < \infty$ *for all responses* $x(t)$ *corresponding to controls of* $\Delta(\mathrm{Lip}\, A)$, *that is,* $|x(t)| < B$ *uniformly for all responses. Then there exists an optimal control* $u^*(t) \in \Delta(\mathrm{Lip}\, A)$, *that is,* $C(u^*) \leqq C(u)$ *for all* $u(t) \in \Delta(\mathrm{Lip}\, A)$.

Proof. Assume that $\Delta(\mathrm{Lip}\, A)$ is infinite, and define

$$\inf C(u) = m > -\infty \quad \text{for all } u(t) \in \Delta(\mathrm{Lip}\, A).$$

Select a sequence $u^{(k)}(t)$ on $t_0^{(k)} \leqq t \leqq t_1^{(k)}$ of control of $\Delta(\mathrm{Lip}\, A)$ with $C(u^{(k)})$ decreasing monotonically towards m as $k = 1, 2, 3, \ldots$ tends to infinity. Select a subsequence (still called $u^{(k)}(t)$) with

$$t_0^{(k)} \to t_0^* \quad \text{and} \quad t_1^{(k)} \to t_1^* \quad \text{monotonically.}$$

Consider first the case where

$$t_0^{(k)} \leqq t_0^* \leqq t_1^* \leqq t_1^{(k)}, \qquad k = 1, 2, 3, \ldots,$$

and again we omit the trivial subcase where $t_0^* = t_1^*$. Using Ascoli's theorem [6], select a subsequence of these controls such that

$$\lim_{k\to\infty} u^{(k)}(t) = u^*(t)$$

uniformly on $t_0^* \leqq t \leqq t_1^*$ and $u^*(t)$ is a continuous function satisfying

$$|u^*(t) - u^*(t')| \leqq A\,|t - t'| \quad \text{for} \quad t_0^* \leqq t, \quad t' \leqq t_1^*.$$

We must show that $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ lies in $\Delta(\mathrm{Lip}\, A)$. The graph $u^*(t) \subset \Omega$, since Ω is compact.

Consider the responses $x^{(k)}(t)$ of $u^{(k)}(t)$ on $t_0^{(k)} \leqq t \leqq t_1^{(k)}$. Here in vector notation

$$x^{(k)}(t) - x^{(p)}(t) = \int_{t_0^{(k)}}^{t} f\big(s, x^{(k)}(s), u^{(k)}(s)\big)\, ds - \int_{t_0^{(p)}}^{t} f\big(s, x^{(p)}(s), u^{(p)}(s)\big)\, ds$$

for $t_0^* \leqq t \leqq t_1^*$ with k and $p > k$ positive integers. Then

$$|x^{(k)}(t) - x^{(p)}(t)| \leqq \int_{t_0^{(k)}}^{t_0^*} |f(s, x^{(k)}(s), u^{(k)}(s))|\, ds + \int_{t_0^{(p)}}^{t_0^*} |f(s, x^{(p)}(s), u^{(p)}(s))|\, ds + $$
$$+ \int_{t_0^*}^{t} |f(s, x^{(k)}(s), u^{(k)}(s)) - f(s, x^{(p)}(s), u^{(p)}(s))|\, ds.$$

Take k so large that $|t_0^* - t_0^{(k)}|$ and $|t_0^* - t_0^{(p)}|$ are very small and $|u^{(k)}(t) - u^{(p)}(t)|$ is uniformly small on $t_0^* \leqq t \leqq t_1^*$. Use the uniform continuity of $f(t, x, u)$ in $\tau_0 \leqq t \leqq \tau_1$, $|x| \leqq B$, $u \in \Omega$, and compute

$$|x^{(k)}(t) - x^{(p)}(t)| \leqq \frac{\varepsilon}{2} + \int_{t_0^*}^{t} |f(s, x^{(k)}(s), u^{(k)}(s)) - f(s, x^{(k)}(s), u^{(p)}(s))|\, ds + $$
$$+ \int_{t_0^*}^{t} |f(s, x^{(k)}(s), u^{(p)}(s)) - f(s, x^{(p)}(s), u^{(p)}(s))|\, ds$$

so

$$|x^{(k)}(t) - x^{(p)}(t)| \leqq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} + \int_{t_0^*}^{t} \left| \frac{\partial f}{\partial x}\left(s, \tilde{x}(s), u^{(p)}(s)\right) \right| \cdot |x^{(k)}(s) - x^{(p)}(s)|\, ds,$$

for an arbitrarily small $\varepsilon > 0$. Write

$$z(t) = |x^{(k)}(t) - x^{(p)}(t)| \quad \text{and} \quad \left| \frac{\partial f}{\partial x}\left(s, \tilde{x}(s), u^{(p)}(s)\right) \right| < \alpha.$$

Then

$$z(t) \leqq \varepsilon + \alpha \int_{t_0^*}^{t} z(s)\, ds.$$

This integral inequality implies

$$z(t) \leqq \varepsilon\, e^{\alpha(t_1^* - t_0^*)}.$$

Therefore CAUCHY'S criterion yields

$$\lim_{k \to \infty} x^{(k)}(t) = x^*(t) \quad \text{uniformly on} \quad t_0^* \leqq t \leqq t_1^*,$$

and $x^*(t)$ is continuous on $t_0^* \leqq t \leqq t_1^*$.

Now

$$x^{(k)}(t) = x_0 + \int_{t_0^{(k)}}^{t_0^*} f\left(s, x^{(k)}(s), u^{(k)}(s)\right) ds + \int_{t_0^*}^{t} f\left(s, x^{(k)}(s), u^{(k)}(s)\right) ds$$

on $t_0^* \leqq t \leqq t_1^*$. Thus

$$x^*(t) = x_0 + \int_{t_0^*}^{t} f\left(s, x^*(s), u^*(s)\right) ds,$$

and $x^*(t)$ is the response to the control $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$. Clearly $x^*(t_0^*) = x_0$. Also

$$x^*(t_1^*) = \lim_{k \to \infty} x^{(k)}(t_1^*) = \lim_{k \to \infty} \left[x^{(k)}(t_1^{(k)}) - x^{(k)}(t_1^{(k)}) + x^{(k)}(t_1^*)\right]$$

so

$$x^*(t_1^*) = \lim_{k \to \infty} x^{(k)}(t_1^{(k)}).$$

Since

$$x^{(k)}(t_1^{(k)}) \in G(t_1^{(k)}),$$

we have

$$x^*(t_1^*) \in G(t_1^*).$$

Thus $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ belongs to $\Delta(\mathrm{Lip}\, A)$. Furthermore

$$C(u^*) = \lim_{k \to \infty} C(u^{(k)}) = m,$$

and $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ is optimal in $\Delta(\mathrm{Lip}\, A)$.

We return to the assumption that $t_0^{(k)} \to t_0^*$ and $t_1^{(k)} \to t_1^*$, monotonically. Suppose we have $t_0^{(k)} \leqq t_0^*$ and $t_1^{(k)} \leqq t_1^*$, for example. Extend each control $u^{(k)}(t)$ to $t_0^{(k)} \leqq t \leqq t_1^*$ by defining $u^{(k)}(t) = u^{(k)}(t_1^{(k)})$ on $t_1^{(k)} \leqq t \leqq t_1^*$. For large k we can also extend the corresponding responses $x^{(k)}(t)$ to the interval $t_0^{(k)} \leqq t \leqq t_1^*$. Again

$$\lim_{k\to\infty} u^{(k)}(t) = u^*(t) \quad \text{uniformly on} \quad t_0^* \leqq t \leqq t_1^*,$$

and $u^*(t)$ satisfies a Lipschitz condition with constant A. As above we show that

$$\lim_{k\to\infty} x^{(k)}(t) = x^*(t) \quad \text{uniformly on} \quad t_0^* \leqq t \leqq t_1^*$$

and that $x^*(t)$ is the response to $u^*(t)$. Also $x^*(t_0^*) = x_0$ and $x^*(t_1^*) \in G(t_1^*)$. Thus $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ belongs to $\Delta(\text{Lip}\,A)$. Again we compute that $u^*(t)$ on $t_0^* \leqq t \leqq t_1^*$ is optimal in $\Delta(\text{Lip}\,A)$. Q.E.D.

Remark. The remarks following Theorem 1 are also applicable here.

Corollary. *Assume the hypotheses of Theorem* 2 *for each Lipschitz constant* $p = 1, 2, 3, \ldots$, *and let* $u_p^*(t)$ *on* $t_0^{(p)} \leqq t \leqq t_1^{(p)}$ *be optimal in* $\Delta(\text{Lip}\,p)$. *In addition assume*

1) the target $G(t)$ *on* $\tau_0 \leqq t \leqq \tau_1$ *is a fixed point* G,

2) for each $s > 0$ *there exists a measurable control* $u_s(t)$ *on* $t_0^{(s)} \leqq t \leqq t_1^{(s)}$ *in* Δ *such that* $C(u_s) \leqq m + s$ *where* $m = \inf C(u)$ *for* $u \in \Delta$ *and also* $t_1^{(s)} \leqq \tau_1 - \varepsilon$ *for some* $\varepsilon > 0$, *independent of* s,

3) there is a uniform bound $|x(t)| < B < \infty$ *for all responses* $x(t)$ *corresponding to measurable controls in* Δ,

4) for each $s > 0$ *there exists a neighborhood* N *of* G *such that each point in* N *can be steered to* G *by a* C^1 *control having a prescribed initial instant* $t_0 < \tau_1 - \varepsilon$, *a duration* $\leqq \varepsilon$, *and a cost* $\leqq s$ (*this condition is considered in Theorem* 4).

Then

$$\lim_{p\to\infty} C(u_p^*) = m = \inf C(u) \quad \textit{for} \quad u \in \Delta.$$

Proof. It is clear that

$$C(u_p^*) \geqq C(u_{p+1}^*) \qquad p = 1, 2, 3, \ldots,$$

so $\lim_{p\to\infty} C(u_p^*)$ exists and is not less than m. But given a measurable control $u_s(t)$ in Δ, there exists a C^∞ control which approximates $u_s(t)$ almost uniformly so that the corresponding response has an end-point in N. But then x_0 can be steered to the target point G by a control $u_A(t)$ in $\Delta(\text{Lip}\,A)$ for some $A > 0$, and with $C(u_A)$ arbitrarily near to m. For $p > A$ we thus have

$$C(u_A) \geqq C(u_p^*),$$

so $\lim_{p\to\infty} C(u_p^*) = m$, as required. Q.E.D.

2. Domain of Controllability

In this section we investigate the nature of hypotheses A) of Theorem 1. For simplicity we consider the problem of steering an initial point to a fixed target point, say the origin. We still maintain that the restraint set Ω is compact, but in this section we require that the zero vector is in the interior of Ω.

Definition. The domain of controllability $\mathscr{C}$ for the differential system

$$\dot{x}^i = f^i(x^1, \dots, x^n, u^1, \dots, u^m) \qquad i = 1, \dots, n$$

for a given restraint set $\Omega \subset R^m$, consists of the set of all points $x_0 \in R^n$ for which there exists a measurable control $u(t) \subset \Omega$, defined on some finite interval, steering x_0 to the origin* of R^n.

Theorem 3. *Consider*

$$\dot{x}^i = f^i(x^1, \dots, x^n; u^1, \dots, u^m), \qquad i = 1, \dots, n,$$

*where*** $f^i(x, u) \in C^1$ *in* $R^n \times \Omega$. *Assume the vector* $u = 0$ *is an interior point of the compact restraint set* $\Omega \subset R^m$. *Assume*

$$f^i(0, 0) = 0, \qquad i = 1, \dots, n.$$

If the matrix $\frac{\partial f^i}{\partial u^j}(0, 0)$ *has rank* n, *then the domain of controllability* $\mathscr{C}$ *is an open connected subset of* R^n *containing the origin.*

Proof. If the origin of R^n is an interior point of $\mathscr{C}$, then it follows from general continuity arguments that $\mathscr{C}$ is open and connected. By continuity there exists a neighborhood N of the origin in Ω such that for any measurable vector $u(t)$ on $-1 \leqq t \leqq 1$ in N the corresponding solution $x(t)$ initiating at $x_0 = 0$ will be defined in R^n for $-1 \leqq t \leqq 1$.

Let

$$\varphi^i(t, u^1, \dots, u^m) \quad \text{on} \quad -1 \leqq t \leqq 1, \qquad i = 1, \dots, n,$$

be the solution of the given differential system for constant $(u^1, \dots, u^m)$ in N and with $\varphi^i(0, u) = 0$. We shall prove that $\frac{\partial \varphi^i}{\partial u^j}(t_0, 0)$, for some $t_0 < 0$, has rank n. Hence the differentiable map

$$u \to x = \varphi(t_0, u)$$

carries a neighborhood of $u_0 = 0$ onto a neighborhood of $x_0 = 0$. Then there is a neighborhood of $x_0 = 0$ which can be steered to the origin in time $-t_0 > 0$, and each required control is just a constant.

Now

$$\frac{\partial \varphi^i}{\partial t} = f^i(\varphi(t, u), u), \qquad \varphi^i(t, 0) = 0.$$

We compute

$$\frac{\partial}{\partial t}\left(\frac{\partial \varphi^i}{\partial u^j}\right) = \frac{\partial f}{\partial x^k}(\varphi, u)\,\frac{\partial \varphi^k}{\partial u^j} + \frac{\partial f^i}{\partial u^j}(\varphi, u).$$

For $u_0 = 0$ write the matrix

$$z(t) = \frac{\partial \varphi}{\partial u}(t, 0).$$

---

* This is sometimes referred to as null controllability.

** That is, $f(x, u) \in C^0$ in $R^n \times \Omega$ and has a C^1 extension to $R^n \times \hat{\Omega}$, where $\hat{\Omega}$ is an open set containing Ω.

Then

$$\frac{dz}{dt} = f_x(0,0)\,z + f_u(0,0), \qquad z(0) = 0.$$

Then

$$z(t) = \left[I + \frac{t}{2!} f_x + \frac{t^2}{3!} f_x^2 + \cdots\right] \qquad [t f_u].$$

For a small $|t_0|$

$$z(t_0) = \frac{\partial \varphi}{\partial u}(t_0, 0)$$

has rank n, since $f_u(0,0)$ has rank n. Q.E.D.

Corollary. *Consider a linear differential system with real constant coefficients*

$$\dot{x}^i = A^i_k x^k + B^i_j u^j \qquad i, k = 1, \ldots, n; \quad j = 1, \ldots, m.$$

Assume the compact restraint set $\Omega \subset R^m$ contains the zero vector in its interior. If the matrix B has rank n, the domain of controllability $\mathscr{C}$ is an open connected set in R^n containing the origin.

A more difficult problem arises when the control has fewer degrees of freedom than the response, *i.e.* when $m < n$. The next theorem considers this case.

Theorem 4. *Consider*

$$\dot{x}^i = f^i(x^1, \ldots, x^n, u^1, \ldots, u^m) \qquad i = 1, \ldots, n$$

where $f(x,u) \in C^1$ in $R^n \times \Omega$. Assume the vector $u = 0$ is an interior point of the compact restraint set $\Omega \subset R^m$. Assume

1) $f^i(0,0) = 0 \qquad i = 1, \ldots, n$

and

2) there exists a vector $v \in R^m$ such that Bv lies in no invariant subspace of A with dimension $\leqq n-1$, where $B = \frac{\partial f}{\partial u}(0,0)$ and $A = \frac{\partial f}{\partial x}(0,0)$ are real matrices.

Then the domain of controllability $\mathscr{C}$ is an open connected subset of R^n containing the origin.

Proof. Consider a neighborhood $N \subset \Omega$ of the origin in R^m and $\tau_1 > 0$ such that for each measurable vector $u(t)$ on $-\tau_1 \leqq t \leqq \tau_1$ with $u(t) \subset N$ the corresponding response $x(t)$ with $x(0) = 0$ is defined on $-\tau_1 \leqq t \leqq \tau_1$.

We shall define a family of controls in N

$$u^j(t, \xi) = u^j(t, \xi_1, \ldots, \xi_n), \qquad j = 1, \ldots, m$$

on $|t| \leqq \tau_1$ and $|\xi| < \varepsilon$ for some $\varepsilon > 0$. Take

$$u(t,\xi) = v\left[X(t, t_1)\,\xi_1 + X\left(t, \frac{t_1}{2}\right)\xi_2 + \cdots + X\left(t, \frac{t_1}{n}\right)\xi_n\right]$$

where v is the vector mentioned in hypothesis 2),

$$X(t,h) = \begin{cases} 1 & \text{on} \quad |t| \leqq h \\ 0 & \text{on} \quad |t| > h, \end{cases}$$

and t_1 is a small number with $0<|t_1|<\tau_1$. For convenience we choose this piecewise continuous function $X(t, h)$, but it would be easy to use a C^∞ function which almost uniformly approximates $X(t, h)$, for each fixed h.

Note $u(t, 0)=0$ and also

$$\frac{\partial u}{\partial \xi^k} = v\,X\left(t, \frac{t_1}{k}\right), \qquad k=1, \ldots, n$$

is differentiable except at $t=t_1/k$.

Let

$$x^i(t, \xi_1, \ldots, \xi_n) = x^i(t, \xi), \qquad i=1, \ldots, n$$

be the response corresponding to $u(t, \xi)$ with

$$x^i(0, \xi) = 0, \qquad i=1, \ldots, n.$$

Note that $x^i(t, 0)=0$, $i=1, \ldots, n$, since $f^i(0, 0)=0$. Now $x(t, \xi)$ is continuous in $(n+1)$ variables. Also, for each fixed t, the map

$$\xi \to x(t, \xi)$$

is a differentiable map of a neighborhood of the origin of R^n into R^n with the origin fixed. If

$$z^i_k(t) = \frac{\partial x^i}{\partial \xi^k}(t, 0) \qquad i, k = 1, \ldots, n$$

is non-singular, for some fixed $t<0$, then the domain of controllability $\mathscr{C}$ is an open connected subset of R^n containing the origin, as required.

Now, in vector notation,

$$x(t, \xi) = \int_0^t f\big(x(s, \xi), u(s, \xi)\big)\, ds,$$

and so

$$\frac{\partial x}{\partial \xi}(t, 0) = \int_0^t \left[A\,\frac{\partial x}{\partial \xi}(s, 0) + B\,\frac{\partial u}{\partial \xi}(s, 0)\right] ds$$

or

$$z(t) = \int_0^t [A\,z(s) + B\,u_\xi(s, 0)]\, ds,$$

so $z(t)$ is continuous near $t=0$. Also dz/dt exists, except at $t=t_1/k$, $k=1, 2, \ldots, n$.

The k^{th} column of the matrix $z(t)$ satisfies

$$z_{(k)}(t) = \int_0^t \left[A\,z_{(k)}(s) + B\,v\,X\left(s, \frac{t_1}{k}\right)\right] ds$$

or $k=1, 2, \ldots, n$, where the subscript $k=1, 2, \ldots, n$ designates the column. Thus we have

$$\frac{d z_{(k)}}{dt} = A\,z_{(k)} + B\,v\,X\left(t, \frac{t_1}{k}\right) \qquad k=1, \ldots, n$$

with $z_{(k)}(0)=0$. But the solution of this linear differential system is

$$z_{(k)}(t) = e^{tA} \int_0^t e^{-sA} B v X\left(x, \frac{t_1}{k}\right) ds$$

or

$$z_{(k)}(t) = e^{tA} \int_0^{t_1/k} e^{-sA} B v \, ds .$$

Thus, for $|t_1| < |t| < \tau_1$,

$$z_{(1)}(t) = e^{tA} \left[t_1 I - \frac{t_1^2}{2!} A + \frac{t_1^3}{3!} A^2 - \frac{t_1^4}{4!} A^3 + \cdots \right] B v ,$$

$$z_{(2)}(t) = e^{tA} \left[\frac{t_1}{2} I - \frac{(t_1/2)^2}{2!} A + \frac{(t_1/2)^3}{3!} A^2 - \cdots \right] B v ,$$

$$\vdots$$

$$z_{(n)}(t) = e^{tA} \left[\frac{t_1}{n} I - \frac{(t_1/n)^2}{2!} A + \frac{(t_1/n)^3}{3!} A^2 - \cdots \right] B v .$$

We shall show the linear independence of the vectors $(e^{-tA})\, z_{(1)}(t), \ldots, (e^{-tA})\, z_{(n)}(t)$.

The Vandermonde determinant

$$\begin{vmatrix} t_1 & t_1^2 & \ldots & t_1^n \\ t_1/2 & (t_1/2)^2 & \ldots & (t_1/2)^n \\ \vdots & \vdots & & \vdots \\ t_1/n & (t_1/n)^2 & \ldots & (t_1/n)^n \end{vmatrix} \neq 0 .$$

Thus it is sufficient to show the linear independence of the vectors

$$\begin{gathered} B v + O(t_1) \\ A B v + O(t_1) \\ A^2 B v + O(t_1) \\ \vdots \\ A^{n-1} B v + O(t_1) , \end{gathered}$$

where $\lim_{t_1 \to 0} O(t_1) = 0$. Since the vectors $Bv, ABv, \ldots, A^{n-1}Bv$ are linearly independent, we conclude that the vectors $z_{(1)}(t), z_{(2)}(t), \ldots, z_{(n)}(t)$ are independent and that the matrix $z(t)$ is non-singular. Q.E.D.

Corollary. *Consider*

$$\dot{x}^i = f^i(x^1, \ldots, x^n, u^1, \ldots, u^m) , \qquad i = 1, \ldots, n$$

where $f(x, u) \in C^1$ *in* $R^n \times \Omega$ *and the origin of* R^m *is an interior point of the compact set* Ω. *Assume*

1) $f^i(0, 0) = 0, \qquad i = 1, \ldots, n,$

2) there exists a vector $v \in R^m$ *such that* Bv *lies in no invariant subspace of* A *with dimension* $\leqq n-1$.

Assume also that

$$\dot{x}^i = f^i(x^1, \ldots, x^n, 0, \ldots, 0) \qquad i = 1, \ldots, n$$

*is globally asymptotically stable** *towards the origin of* R^n. *Then the domain of controllability* $\mathscr{C}$ *is the entire space* R^n.

---

* See reference [*12*].

Corollary. *Consider*

$$x^{(n)}+a_{n-1}(x, x', \ldots, x^{(n-1)})\, x^{(n-1)}+\cdots+a_0(x, x', \ldots, x^{(n-1)})\, x=u$$

where the coefficients $a_i(x_1, x_2, \ldots, x_n) \in C^1$ *in* R^n *and* $|u| \leqq \varepsilon$, *for some* $\varepsilon > 0$. *We write the corresponding first order system*

$$S) \qquad \begin{cases} \dot{x}_1 = x_2, \\ \dot{x}_2 = x_3, \\ \quad \vdots \\ \dot{x}_n = -a_{n-1}(x_1, \ldots, x_n)\, x_n - \cdots - a_0(x_1, \ldots, x_n)\, x_1 + u. \end{cases}$$

If the system

$$\begin{aligned} \dot{x}_1 &= x_2, \\ &\;\vdots \\ \dot{x}_n &= -a_{n-1}(x_1, \ldots, x_n)\, x_n - \cdots - a_0(x_1, \ldots, x_n)\, x_1 \end{aligned}$$

is globally asymptotically stable towards the origin, then for S) hypotheses 1) and 2) of the theorem automatically hold and the domain of controllability for S) is all R^n.

Corollary. *Consider*

$$\dot{x}^i = f^i(x^1, \ldots, x^n, u^1, \ldots, u^m), \qquad i = 1, 2, \ldots, n$$

where $f(x, u) \in C^1$ *for* $x \in R^n$ *and* $u \in U \subset R^m$. *Assume that for each point* $x_0 \in R^n$ *there exists an interior point* $u(x_0)$ *of U such that*

1) $f^i(x_0, u(x_0)) = 0, \qquad i = 1, \ldots, n$

and

2) there exists a vector $v \in R^m$ *such that* Bv *lies in no invariant subspace of A with dimension* $\leqq n-1$, *where* $B = \frac{\partial f}{\partial u}(x_0, u(x_0))$ *and* $A = \frac{\partial f}{\partial x}(x_0, u(x_0))$ *are real matrices. Then for each pair of points* x_1 *and* x_2 *of* R^n *there exists a piecewise continuous controller* $u(t)$ *on* $t_1 \leqq t \leqq t_2$ *in U which steers the response* $x(t)$ *from* $x(t_1) = x_1$ *to* $x(t_2) = x_2$.

Proof. In the proof of Theorem 4 we note that there is a neighborhood of the origin which consists of points that can be steered to the origin by a piecewise continuous controller in a finite time. Also there is a neighborhood of the origin consisting of points to which the origin can be steered by a piecewise continuous controller in a finite time.

The same properties hold for each point $\bar{x}$ in R^n, after translating both x and u as required.

Now let S be the subset of R^n consisting of points to which x_1 can be steered by a piecewise continuous controller in a finite time interval. Clearly S is both open and closed in R^n. Thus $S = R^n$ and x_1 can be steered to x_2, as required, *cf.* [7]. Q.E.D.

Remarks. In the previous corollaries we can replace the hypothesis that $\dot{x}^i = f^i(x^1, \ldots, x^n, u^1, \ldots, u^m)$ be asymptotically stable for $u^j \equiv 0$, $j = 1, \ldots, m$, by the hypothesis that $\dot{x}^i = f^i(x^1, \ldots, x^n, u^1, \ldots, u^m)$ be asymptotically stable for some appropriate choice of $u^j = u^j(x^1, \ldots, x^n)$. As a particular choice of $u^j = u^j(x)$ to satisfy the condition of asymptotic stability consider a Lyapunov function $V(x) \geqq 0$, $V(x) \in C^1$, with the choice of $u \in \Omega$ which minimizes $\left[\frac{\partial V}{\partial x^i} f^i(x, u)\right]$. If this minimum is negative definite in $0 < |x| < \infty$ for some $V(x) > 0$, with $V(x) \to \infty$ as $|x| \to \infty$, then the domain of controllability $\mathscr{C}$ is all of R^n.

We next consider the domain of controllability for certain second order differential equations

$$\ddot{x} + f(x, \dot{x}, u) = 0$$

which we write as the first order system

$$\dot{x} = y, \quad \dot{y} = -f(x, y, u)$$

in the phase plane.

Theorem 5. *Consider the differential equation*

$$\ddot{x} + f(x, \dot{x}, u) = 0$$

with $f(x, y, u) \in C^1$ *for all* (x, y) *and* $u \in \Omega$, *where* Ω *is a compact interval containing zero as an interior point. Assume*

a) $f(0, 0, 0) = 0$

b) $f_x(x, y, 0) > 0$ *and* $f_y(x, y, 0) > 0$ *everywhere*

c) $f_u(0, 0, 0) \neq 0$.

Then the domain of controllability $\mathscr{C}$ *is the entire* (x, y)*-plane.*

Proof. The system

$$\dot{x} = y$$
$$\dot{y} = -f(x, y, 0)$$

has the variable Jacobian matrix $J = \begin{pmatrix} 0 & 1 \\ -f_x & -f_y \end{pmatrix}$. Since the eigenvalues of J have negative real parts for all (x, y), the system is globally asymptotically stable to the origin, *cf.* [*12*]. However the solutions will not reach the origin in a finite time.

Now

$$A = \begin{pmatrix} 0 & 1 \\ -f_x(0, 0, 0) & -f_y(0, 0, 0) \end{pmatrix}$$

and

$$B = \begin{pmatrix} 0 \\ -f_u(0, 0, 0) \end{pmatrix}.$$

Take $v = 1$ as a 1-vector; then

$$B v = \begin{pmatrix} 0 \\ -f_u(0, 0, 0) \end{pmatrix} \quad \text{and} \quad A B v = \begin{pmatrix} -f_u(0, 0, 0) \\ f_u \cdot f_y \end{pmatrix}$$

are linearly independent. Thus the domain of controllability to the origin is an open set. Therefore $\mathscr{C}$ is the entire (x, y)-plane. Q.E.D.

Corollary. *Consider* $$\ddot{x} + f(\dot{x}) + g(x) = u$$

where $f(y)$ *and* $g(x) \in C^1$ *for all* (x, y) *and* $|u| \leqq \varepsilon$, *for some* $\varepsilon > 0$. *Assume*

a) $f(0) = g(0) = 0$

b) $f'(y) > 0$ *and* $g'(x) > 0$ *everywhere.*

Then the domain of controllability $\mathscr{C}$ *is the entire* (x, y)*-plane.*

We now turn from the asymptotically stable case, where the domain of controllability $\mathscr{C}$ is the entire space, to the conditionally unstable saddle-point case, where $\mathscr{C}$ is a strip in the plane.

Lemma 1. *Consider*

$$\mathscr{D}) \qquad \ddot{x} + f(\dot{x}) - g(x) = 0$$

with $f(y)$, $g(x) \in C^1$ *and* $f(0) = g(0) = 0$, $f'(y) > 0$, $g'(x) > 0$ *for all* (x, y) *in* R^2. *Then the solution curve family of* $\mathscr{D}$) *in the* $(x, y = \dot{x})$ *phase plane is topologically equivalent to the solution curve family of the linear equation*

$$\ddot{x} + \dot{x} - x = 0.$$

Proof. Write the equation $\mathscr{D}$) in the phase plane as

$$\mathscr{S}) \qquad \begin{cases} \dot{x} = y, \\ \dot{y} = g(x) - f(y). \end{cases}$$

We seek a homeomorphism of R^2 onto R^2 which carries the solution curves (sensed but not parametrized) of $\mathscr{S}$) onto those of

$$\mathscr{L}) \qquad \begin{cases} \dot{x} = y, \\ \dot{y} = x - y. \end{cases}$$

In $\mathscr{S}$) the origin is the unique critical point. The variational equation o $\mathscr{S}$) at the origin is

$$\begin{cases} \dot{x} = y, \\ \dot{y} = g'(0)\, x - f'(0)\, y, \end{cases}$$

with real eigenvalues of opposite signs. Thus $\mathscr{S}$) and $\mathscr{L}$) are topologically equivalent near the origin.

Thus $\mathscr{S}$) has exactly four solution curves which have the origin as a limit point and these approach the origin with the same directions as for $\mathscr{L}$), that is, one of these solutions of $\mathscr{S}$) lies, near the origin, in each of the four quadrants. Call these solutions of $\mathscr{S}$), which will be shown to be separatrices in the sense of MARKUS [*13*], by the numerals I, II, III, IV corresponding to the quadrants in which they approach the origin. By examining $\mathscr{S}$) on each of the coordinate axes it is easy to see that each of I, II, III, IV lies entirely in the corresponding quadrant.

Along I we have $$\frac{dy}{dx} = \frac{g(x) - f(y)}{y}$$

which is bounded from above on each compact x-interval. Hence I is single-valued over the entire positive x-axis. A similar argument shows that III is single-valued over the entire negative x-axis. The separatrices II and IV are each single-valued over a segment of the x-axis, and they extend to infinity, that is, $x^2+y^2\to\infty$ on II and IV.

Consider a solution curve of $\mathscr{S})$ in the plane sector bounded by separatrices II and I. Here $\dot{x}=y>0$ and, in the second quadrant $\dot{y}<0$. Thus each solution curve in this sector must intersect the first quadrant and must exist in the first quadrant for all larger values of x to the right of any such intersection point. Hence the solution curves in the sector between II and I are linearly ordered (using the inclusion relation for the regions above the curves), and they form a parallel family, as described by MARKUS [*13*]. A similar analysis holds for the sector between III and IV.

In the sector bounded by IV and I each solution of $\mathscr{S})$ intersects the positive x-axis, which thereby serves as a transversal for the solutions in this sector. A similar analysis holds in the sector bounded by II and III.

Thus the only separatrices of $\mathscr{S})$ are the critical point at the origin and the solution curves I, II, III, and IV. Therefore by the general theory of separatrix configurations [*13*] we find that $\mathscr{S})$ is homeomorphic with $\mathscr{L})$, as required. Q.E.D.

Remark. Call the solution curves II and IV, leading towards the critical point of $\mathscr{S})$ with negative slope, the principal separatrices, and call I and III the minor separatrices of $\mathscr{S})$. The principal separatrices of $\mathscr{S})$, together with the critical point at the origin, form a topological image of a line in R^2 which is single-valued over the entire y-axis. The same holds for the minor separatrices of $\mathscr{S})$ over the entire x-axis.

Lemma 2. *Consider the two differential systems*

$$\mathscr{S}_{\pm}) \qquad \begin{cases} \dot{x}=y, \\ \dot{y}=g_{\pm}(x)-f(y) \end{cases}$$

where $f(y)$ *and* $g_{\pm}(x)=g(x)\pm C_{\pm}$, *for positive constants* $C_{\pm}$, *are in* C^1 *in* R^2. *Assume* $f(0)=g(0)=0$, $f'(y)>0$, $g'(x)>0$ *everywhere in* R^2, *and* $|g(x)|\to\infty$ *as* $|x|\to\infty$.

Then each of $\mathscr{S}_{+})$ *and* $\mathscr{S}_{-})$ *is topologically equivalent to the linear system, in the phase plane* $(x, y=\dot{x})$

$$\mathscr{L}) \qquad \begin{cases} \dot{x}=y, \\ \dot{y}=x-y. \end{cases}$$

The principal separatrices $\mathrm{II}_{\pm}$ *and* $\mathrm{IV}_{\pm}$, *together with the critical point, of* $\mathscr{S}_{\pm})$ *form the topological image of a line which separates the plane in two. At each fixed ordinate the principal separatrix or critical point of* $\mathscr{S}_{-})$ *lies to the right of the corresponding principal separatrix or critical point of* $\mathscr{S}_{+})$. *The open band* B *between the principal separatrices and critical points of* $\mathscr{S}_{+})$ *and* $\mathscr{S}_{-})$ *is homeomorphic to an open band between parallel straight lines in the plane. Also each minor separatrix of* $\mathscr{S}_{\pm})$ *intersects the principal separatrices of* $\mathscr{S}_{\mp})$ *in at most one point, respectively (see Figure 1).*

Proof. By Lemma 1 the topological configuration of the solutions of $\mathscr{S}_{+})$ or of $\mathscr{S}_{-})$ is that of $\mathscr{L})$. It is clear that the critical point of $\mathscr{S}_{-})$ lies to the right of the critical point of $\mathscr{S}_{+})$, on the x-axis. By the preceeding remark the principal separatrices II_- and IV_-, together with the critical point of $\mathscr{S}_-)$, form the topological image of a line which is a single-valued covering of the y-axis. The same holds for $\mathscr{S}_+)$.

Suppose II_+ and II_- intersect. At the intersection point which is furthest to the right the slope of II_- is less than that of II_+, and this contradicts the supposition of an intersection. Similarly IV_+ and IV_- cannot intersect.

Therefore the open band B bounded by the principal separatrices and critical points of $\mathscr{S}_+)$ and $\mathscr{S}_-)$ is homeomorphic to the open band between two parallel straight lines in the plane. In fact, the homeomorphism can be extended to the closed bands.

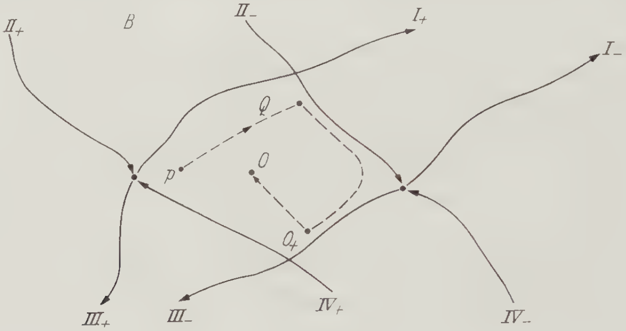

Fig. 1

Now on a minor separatrix, say I_+, we have $\dot{x} > 0$ and $\dot{y} > 0$. Hence I_+ intersects II_- in exactly one point. Similarly III_- intersects IV_+ in exactly one point. Q.E.D.

Theorem 6. *Consider*

$$\mathscr{D})\qquad \ddot{x} + f(\dot{x}) - g(x) = u,$$

where $f(y), g(x) \in C^1$ *for all* (x, y) *in* R^2 *and* $-C_- \leqq u \leqq C_+$ *is the finite real interval* Ω *with* $-C_- < 0 < C_+$. *Assume* $f(0) = g(0) = 0$, $f'(y) > 0$, $g'(x) > 0$ *in* R^2 *and* $|g(x)| \to \infty$ *as* $|x| \to \infty$. *Then the domain of controllability* $\mathscr{C}$ *in the phase plane* $(x, y = \dot{x})$ *is precisely the open topological band* B *bounded by the principal separatrices, and critical points, of the two systems*

$$\mathscr{S}_\pm)\qquad \begin{cases} \dot{x} = y, \\ \dot{y} = g(x) \pm (C_\pm) - f(y). \end{cases}$$

Proof. Consider the open quadrilateral Q in R^2 bounded by I_+, IV_+ and II_-, III_-. The origin 0 lies in Q, and the solution $\mathscr{S}_+)$ through 0 must pass through a point 0_+ in Q very near III_- before it comes to 0.

Now take an initial point p in Q. Steer p by the solution of $\mathscr{S}_+)$ until it almost reaches the intersection of this solution with II_-. Then follow the solution of $\mathscr{S}_-)$ around the inside of the boundary of Q to the point 0_+. Then switch to the solution of $\mathscr{S}_+)$ which steers the point to 0.

If p lies in the band B but above the quadrilateral Q, start out along a solution of $\mathscr{S}_+)$ until one almost reaches the intersection with II_-. Then switch to a solution of $\mathscr{S}_-)$ and follow this until the point 0_+ is again reached. Then follow the solution of $\mathscr{S}_+)$ into 0.

If p lies in the band B but below the quadrilateral Q, start out along a solution of $\mathscr{S}_-)$ until one almost reaches IV_+. Then follow $\mathscr{S}_+)$ into Q, and proceed as above.

Now let Z be a point of $R^2 - B$. If Z lies in $y > 0$ and to the right of the critical point of $\mathscr{S}_-)$, then Z cannot be steered out of this quadrant by a controller $-C_- \leqq u(t) \leqq C_+$ and thus Z cannot be steered to 0. Now Z follows a solution curve of

$$\begin{cases} \dot{x} = y, \\ \dot{y} = g_-(x) - f(y) + \varepsilon(t) \end{cases}$$

where $\varepsilon(t) = u(t) + C_- \geqq 0$ is measurable on some finite time interval. Thus if Z lies in the open sector bounded by II_- and I_-, then Z must lie above a solution curve of $\mathscr{S}_-)$, and hence, using the parallel structure of the solution curve family in this sector, we see that Z cannot be steered to 0. If Z lies in the open sector bounded by IV_- and I_-, then Z must enter the half-plane $y > 0$ to the right of the critical point of $\mathscr{S}_-)$, whatever the controller $u(t)$. Therefore, if Z lies to the right of the closed band $\bar{B}$, then Z cannot be steered to the origin 0 by a measurable controller $u(t)$ with $-C_- \leqq u(t) \leqq C_+$. A similar argument shows that if Z lies to the right of the closed band B, it cannot be steered to the origin 0.

By Theorem 4 the domain of controllability $\mathscr{C}$ is an open plane set. Therefore $\mathscr{C} = B$. Q.E.D.

Remark. Each initial point in $\mathscr{C} = B$ can be steered to the origin 0 using only the solutions of $\mathscr{S}_+)$ and $\mathscr{S}_-)$. It is necessary to use only two switches between these two systems for each point in $\mathscr{C}$ (even though the discussion in the theorem uses three switches in certain cases).

As a final example consider the van der Pol equation

$$\ddot{x} + \mu(x^2 - 1)\,\dot{x} + x = 0$$

for a positive constant μ. In the phase plane this is the system

$$\begin{cases} \dot{x} = y \\ \dot{y} = \mu(1 - x^2)\,y - x. \end{cases}$$

The origin 0 is the unique critical point and is an unstable focus, or node. There is a unique periodic solutions S_μ which is orbitally asymptotically stable and which lies between the abscissas $x = \pm d_\mu$ (for estimates of d_μ see [*14*]).

Theorem 7. *Consider the differential system*

$$\mathscr{S}_\mu) \qquad \begin{cases} \dot{x} = y, \\ \dot{y} = \mu[1 - (x - u)^2]\,y - (x - u) \end{cases}$$

for a fixed $\mu > 0$ *and measurable controls* $u(t)$ *defined on finite intervals with* $|u(t)| \leqq \varepsilon$, *for some* $\varepsilon > 0$. *Then the domain of controllability* $\mathscr{C}$ *to the origin* 0 *is all* R^2 *whenever*

$$\varepsilon > d_\mu/2.$$

Proof. For each choice of a constant u on $|u| \leqq \varepsilon$ the corresponding system $\mathcal{S}_\mu)$ has a unique critical point 0_μ, which is unstable, and a unique periodic solution, which is orbitally asymptotically stable and which lies between the abscissas $x = u \pm d_\mu$.

If the constant u satisfies $|u| < d_\mu$, then it is easy to see that 0_μ belongs to $\mathscr{C}$. If $|u| < d_\mu$ and if $\mathscr{C}$ intersects the periodic solution S_μ, then $\mathscr{C} = R^2$.

Now if $\varepsilon > d_\mu/2$, we can choose a constant u so $|u| < \varepsilon$ and yet $|2u| > d_\mu$. Then $\mathscr{C}$ intersects S_μ, and hence $\mathscr{C} = R^2$. Q.E.D.

The work of L. MARKUS was supported by NSF grant 11287 and OOR contract DA-11-022-ORD-3369.

References

- [1] ALEXANDROFF, P., & H. HOPF: Topology. Berlin: Springer 1935.
- [2] BELLMAN, R., I. GLICKSBERG & O. GROSS: On the bang-bang control problem. Quart. J. Appl. Math. **14**, 11—18 (1956).
- [3] BOLTANSKII, V. G., R. V. GAMKRELIDZE & L. S. PONTRIAGIN: The theory of optimal processes (I — The maximum principle). Izvestiya Akademii Nauk. SSSR. Seriya Mathematicheskaya **24**, 3—42 (1960).
- [4] CODDINGTON, E. A., & N. L. LEVINSON: Theory of Ordinary Differential Equations. New York: McGraw-Hill Book Co. 1955.
- [5] DUNFORD, N., & J. SCHWARTZ: Linear Operators. New York: Interscience Publishers 1959.
- [6] GRAVES, L. M.: The Theory of Functions of Real Variables. New York: McGraw-Hill Book Company 1946.
- [7] KALMAN, R. E.: On the general theory of control systems. International Federation of Automatic Control Congress **4**, 2020—2030 (1960).
- [8] KRASOVSKII, N. N.: On a problem of optimum control of nonlinear systems. P.M.M. **23**, No. 2, 209—229 (1959).
- [9] LA SALLE, J. P.: Time optimal control systems. Proc. Nat. Acad. Sci. **45**, 573—577 (1959).
- [10] LEE, E. B.: Methods of Optimum Feedback Control. University of Minnesota thesis, Minneapolis, Minnesota 1960.
- [11] MARKUS, L., & E. B. LEE: On the existence of optimal controls. Trans. ASME J. of Basic Engineering (to appear).
- [12] MARKUS, L., & H. YAMABE: Global stability criteria for differential systems. Osaka J. of Math. 1961.
- [13] MARKUS, L.: Global structure of ordinary differential equations in the plane. Trans. Amer. Math. Soc. **76**, No. 1, 127—148 (1954).
- [14] URABE, M., H. YANAGIARA & Y. SHINOHARA: Periodic solutions of van der Pol's equation with damping coefficient 2∼10. J. of Science of Hiroshima University 1960.

Mathematics Department
Institute of Technology
University of Minnesota
Minneapolis

(Received March 6, 1961)

Finite Dimensionality of the Schrödinger Operator Bottom

F. H. BROWNELL

Communicated by M. M. SCHIFFER

1. Introduction

We consider in the complex Hilbert space $L_2(E_n)$, E_n being n-dimensional Euclidean space with $n \geqq 1$, the Schrödinger partial differential operator H_0 defined by

$$[H_0 u](\boldsymbol{x}) = -\nabla^2 u(\boldsymbol{x}) + V(\boldsymbol{x})\, u(\boldsymbol{x}) \tag{1.1}$$

over $\boldsymbol{x} \in E_n$ for $u \in \mathscr{D}_0$, the domain $\mathscr{D}_0$ of H_0 being the set of all complex-valued functions u over E_n such that $Vu \in L_2(E_n)$, that all partial derivatives of u of order $\leqq 2$ exist and are continuous throughout E_n, and that u and all such partials remain bounded over E_n after multiplication by $(1+|\boldsymbol{x}|^m)^{-1}\exp(\frac{1}{2}|\boldsymbol{x}|^2)$ for some $m \geqq 0$ depending on u. Here $|\boldsymbol{x}| = \left[\sum_{j=1}^{n} |x_j|^2\right]^{\frac{1}{2}}$, $\nabla^2 = \sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2}$ is the Laplacian, and V is a fixed, measurable, real-valued function over E_n subject to conditions stated in detail in the next section. These conditions suffice to specify a self-adjoint extension H of H_0. Let E be the spectral measure of H, and let $N(\lambda)$ be the dimension of the $(-\infty, \lambda]$ eigenspace $E((-\infty, \lambda])\, L_2(E_n)$ for real λ. Our main result is the proof (under our conditions on V, which imply that $\lim_{r\to+\infty} (\operatorname{ess\,inf}_{|\boldsymbol{x}|\geqq r} V(\boldsymbol{x})) \geqq c$) that $N(\lambda) \leqq N_0 < +\infty$ over all $\lambda < c$. By adding a suitable constant, we may and hereafter do consider only the $c=0$ case for simplicity. In the next section we state precisely the conditions on V and our resulting main theorem. We remark that these conditions are implied by the more stringent ones of IKEBE [*6*], for which our conclusion here thus holds. Actually the variational machinery developed in a previous paper on asymptotics [*3*] enables us readily to reduce our theorem to some well known lemmas about ordinary differential equations (also stated in the next section). The stated result for (1.1) from these known tools does seem to be new.

2. Hypotheses and Main Statements

Let μ_n denote n-dimensional Lebesgue measure for E_n, and let "measurable" mean measurable with respect to the σ-algebra of E_n subsets which is the domain of μ_n. With this notation our hypotheses about V are stated as follows. $L_2(D)$

This research was supported financially by the National Science Foundation, grant NSF-G 11097.

for measurable $D \subseteqq E_n$ means the complex L_2-space formed with respect to μ_n-integration over D.

Condition 2.A. *Let V be a real-valued, measurable function over E_n which is essentially bounded over $\{\boldsymbol{x} \mid |\boldsymbol{x}| \geqq R_0\}$ for some real $R_0 > 0$ and which for some real $\gamma \geqq 1$ having $\gamma > \frac{1}{2} n$ satisfies*

$$\int_{|\boldsymbol{x}| \leqq R_0} |V(\boldsymbol{x})|^{\gamma} \, d\mu_n(\boldsymbol{x}) < +\infty. \tag{2.1}$$

Condition 2.B. *Let V satisfy Condition 2.A, and in addition let $V(\boldsymbol{x}) \geqq -g(|\boldsymbol{x}|)$ almost (μ_n) everywhere over $|\boldsymbol{x}| \geqq R_0$, where $g(\mathrm{r})$ is defined and real-valued and possesses a continuous derivative over real $\mathrm{r} \geqq R_0$ with $g(\mathrm{r}) \geqq 0$ there such that, for some finite M and $R \geqq R_0$:*

$$g(\mathrm{r}) \leqq M \mathrm{r}^{-2} \qquad \textit{over} \quad \mathrm{r} \in [R_0, +\infty); \tag{2.2}$$

$$\mathrm{r} \int_{\mathrm{r}}^{\infty} g(s) \, ds \leqq \tfrac{1}{4} \qquad \textit{over} \quad \mathrm{r} \in [R, +\infty) \quad \textit{if} \quad |n-2| \geqq 1; \tag{2.3}$$

$$\mathrm{r} \int_{\mathrm{r}}^{\infty} g(s) \, ds \leqq \tfrac{1}{4} \mathrm{r} \int_{\mathrm{r}}^{\infty} s^{-1} (\ln s)^{-2} \, ds \qquad \textit{over} \quad [R, +\infty) \qquad \textit{if} \quad n = 2. \tag{2.3)'$$

Notice that both (2.2) and (2.3) are implied if $1 > \limsup_{\mathrm{r} \to -\infty} 4 \mathrm{r}^2 g(\mathrm{r})$; likewise both (2.2) and (2.3)′ are implied if $1 > \limsup_{\mathrm{r} \to +\infty} 4 \mathrm{r}^2 (\ln \mathrm{r})^2 g(\mathrm{r})$. Also (2.3) is implied by $\int_{R_0}^{\infty} \mathrm{r} \, g(\mathrm{r}) \, d\mathrm{r} < +\infty$, and (2.3)′ is implied by $\int_{R_0}^{\infty} \mathrm{r} (\ln \mathrm{r})^2 g(\mathrm{r}) \, d\mathrm{r} < +\infty$, since the right side of (2.3)′ has the asymptotic form $\frac{1}{4} (\ln \mathrm{r}) \; {}^2 \{1 + O((\ln \mathrm{r})^{-1})\}$ as $\mathrm{r} \to +\infty$.

Assuming only Condition 2.A for this purpose, we now specify our self-adjoint extension H of H_0 in $L_2(E_n)$ as follows. Define $H = G_{\omega}^{-1} - \omega^2 I$ as in Theorem T.5 of [2] (see T.3 and T.4 there also, pp. 959—961), H being self-adjoint since G_{ω} is a bounded Hermitian integral operator on $L_2(E_n)$ with Green function kernel. Also $H_0 \subseteqq H$ as desired for extension ([2], Theorem T.6, p. 961 except for the trivially justified replacement of $\check{\mathscr{D}}_1$ there by $\mathscr{D}_0$ of (1.1) here). Other than these references to [2], we here give no further details of this construction of H, since for $\gamma \geqq 2$ in Condition 2.A, as is required if dimension $n \geqq 4$, this construction will now be seen unnecessary to specify H.

Now assume Condition 2.A with $\gamma \geqq 2$ also. Then the condition $V u \in L_2$ in the definition of $\mathscr{D}_0$ following (1.1) may be dropped as redundant, and $\mathscr{D}_0$ is dense in $L_2(E_n)$ and H_0 is symmetric. Also under Condition 2.A with $\gamma \geqq 2$ there exists a unique self-adjoint extension H' of H_0 (by [2], Theorem T.1, p. 954 with $H_1 \subseteqq H_0 \subseteqq H'$ and H' the unique self-adjoint extension of H_1 and thus also of H_0). Thus the self-adjoint extension H constructed above must coincide with H', and H here may be specified as this unique self-adjoint extension of H_0.

Thus we have specified our self-adjoint extension H of H_0, using Condition 2.A only. With $N(\lambda)$ defined in terms of the spectral measure of H as in the Introduction, we may state our main theorem as follows.

Theorem 2.C. *Let V satisfy both Conditions* 2.A *and* 2.B. *Then $N(\lambda) \leqq N_0$ over real $\lambda < 0$ for some positive integer $N_0 < +\infty$.*

By using the variational machinery in [*3*], we can reduce the proof of Theorem 2.C to some well known lemmas about ordinary differential equations, this reduction process making no use of either (2.2) or (2.3) except for the condition that $0 = \lim_{r \to +\infty} g(r)$, as is obviously implied by (2.2). In Section 3 we shall describe this reduction procedure and in Section 4 discusss and give the references for these lemmas from which Conditions 2.3 and 2.3′ arose.

For these lemmas we consider for real R the differential equation over $(R, +\infty)$

$$-u''(r) - f(r)\,u(r) = -\omega^2 u(r), \tag{2.4}$$

where by a solution u of (2.4) over $(R, +\infty)$ we mean a complex-valued function there such that $u''(r)$ exists satisfying (2.4) at each real $r > R$. We shall be interested also in solutions u of (2.4) which satisfy at R a boundary condition of the form

$$0 = (\cos\vartheta)\,u'(R^+) - (\sin\vartheta)\,u(R^+) \tag{2.5}$$

for a given real ϑ, where the right-hand limits $u'(R^+)$ and $u(R^+)$ are assumed to exist and be finite.

Lemma 2.D. *Let the function f be continuous and real-valued over $[R, +\infty)$ and satisfy $f(r) \geqq 0$ there for some real $R_1 > 0$. In addition let f satisfy either*

$$r \int_r^\infty f(s)\,ds \leqq \tfrac{1}{4} \tag{2.6}$$

over $r \geqq R_1$ or else satisfy $f(r) = \frac{1}{4} r^{-2} + \tilde{f}(r)$ with

$$r \int_r^\infty \tilde{f}(s)\,ds \leqq \tfrac{1}{4}\, r \int_r^\infty s^{-2} (\ln s)^{-2}\,ds \tag{2.6$'$}$$

over $r \geqq R_1$ (the Lebesgue integral on the left of (2.6)′ existing since $\tilde{f}(s) \geqq -\frac{1}{4} s^{-2}$). Then there exists some real $R' \geqq R_1$ such that for every real ϑ and real $R \geqq R'$ there exists at most a single real $\omega > 0$ having a solution u of (2.4) and (2.5) for R which also satisfies $0 < \int_R^\infty |u(r)|^2\,dr < +\infty$.

Finally we record the following well known and trivial lemma (differing only in boundary condition from Theorem 3.9, [*3*]). Solutions are defined as for (2.4).

Lemma 2.E. *Let q be a real-valued, bounded, continuous function over $(R, +\infty)$ for some real R, and let the real number λ satisfy*

$$\lambda < \inf_{r > R} q(r). \tag{2.7}$$

Then for each real $k \geqq 0$ there exists no real-valued solution u of

$$-u''(r) + q(r)\,u(r) = \lambda\,u(r) \tag{2.8}$$

over $(R, +\infty)$ which also satisfies $0 < \int_R^\infty |u|^2\,d\mu_1 < +\infty$ and has finite right-hand limits $u'(R^+)$ and $u(R^+)$ satisfying

$$0 = u'(R^+) - k\,u(R^+). \tag{2.9}$$

In closing this section, we remark that Theorem 2.C taken together with our asymptotic formula in [*3*] indicates, by comparison of (2.2) here with (2.1) to (2.6) in [*3*], that $O(\mathrm{r}^{-2})$ at ∞ for the potential function is roughly speaking the borderline between $N(\lambda)$ bounded or unbounded as $\lambda \to 0^-$. In addition to this description of the $(-\infty, 0)$ part of the spectrum of H, theorems describing the $[0, +\infty)$ part of the spectrum are also known ([*1*], Theorem 6.1, p. 573, Theorem 6.7, p. 586, Theorem 6.8, p. 591; [*6*], Theorem 6, p. 20; [*7*], Theorem 1 and 1a, pp. 405—406).

3. Reduction Proof

In this section we shall prove Theorem 2.C, assuming valid Lemmas 2.D and 2.E. These lemmas will be proved in the next section. Thus here we assume Conditions 2.A and 2.B to be satisfied; actually, as remarked in Section 2, until treating the ordinary differential equations at the end of this proof, we shall make no use of (2.2) or (2.3) here except for the fact that $\lim_{\mathrm{r} \to +\infty} g(\mathrm{r}) = 0$, obviously implied by (2.2).

Now to employ the variational machinery ([*3*], Section 3 plus Addendum) needed here, we consider $D_b^+ = \{\boldsymbol{x} \in E_n \mid b < |\boldsymbol{x}|\}$ and $D_b^- = \{\boldsymbol{x} \in E_n \mid b > |\boldsymbol{x}|\}$ for $b \in (0, +\infty)$, and with $D = D_b^\pm$ or $D = E_n$ we consider $\Omega_D \subseteqq L_2(D)$ defined as follows. Ω_D is the set of all $u \in L_2(D)$ such that u possesses throughout D continuous partials of all orders $\leqq 2$, that all such partials possess continuous extensions over the closure $\overline{D}$ of D, and (if D is unbounded) that u and all such partials are each in absolute value $\leqq M(1 + |\boldsymbol{x}|^\alpha) \exp(-2^{-1}|\boldsymbol{x}|^2)$ over $\boldsymbol{x} \in D$ for some M and α in $(0, +\infty)$ depending on u. Finally, with the function $q(\boldsymbol{x})$ defined either by $q(\boldsymbol{x}) = V(\boldsymbol{x})$ or $q(\boldsymbol{x}) = -g(|\boldsymbol{x}|)$ in Conditions 2.A and 2.B, we define for such D and q and for each integer $p \geqq 1$

$$\tau_p(D, q) = \sup_S \left[\inf_u \int_D \left\{ |\boldsymbol{\nabla} u(\boldsymbol{x})|^2 + q(\boldsymbol{x}) \, |u(\boldsymbol{x})|^2 \right\} d\mu_n(\boldsymbol{x}) \right], \tag{3.1}$$

where $|\boldsymbol{\nabla} u(\boldsymbol{x})| = \sum_{j=1}^{n} \left| \frac{\partial}{\partial x_j} u(\boldsymbol{x}) \right|^2$, where the outside supremum is taken over all sets $S \subseteqq L_2(D)$ having card $S = p - 1$ (and thus $S = \emptyset$, the empty set, if $p = 1$), and where the inside infimum is taken over all $u \in \Omega_D$ such that $\int_D u \overline{w} \, d\mu_n = 0$ for each $w \in S$ and $\int_D |u|^2 \, d\mu_n = 1$. Finally, define the non-negative integer $m(\lambda; D, q)$, for real λ and D and q as in (3.1), as the number of integers $p \geqq 1$ such that $\tau_p(D, q) \leqq \lambda$.

With this notation, Conditions 2.A and 2.B together imply for $b - 1 \geqq R_0$ the following inequalities over $\lambda < 0$:

$$N(\lambda) = m(\lambda; E_n, V) < +\infty, \tag{3.2}$$

$$m(\lambda; E_n, V) \leqq m(\lambda; D_b^-, V) + m(\lambda; D_b^+, V), \tag{3.3}$$

$$m(\lambda; D_b^-, V) \leqq m(0; D_b^-, V), \tag{3.4}$$

$$m(0; D_b^-, V) < +\infty, \tag{3.5}$$

$$m(\lambda; D_b^+, V) \leqq m(\lambda; D_b^+, q_0), \tag{3.6}$$

where $q_0(\boldsymbol{x}) = -g(|\boldsymbol{x}|)$ is defined over $\boldsymbol{x} \in D_b^+$.

Here (3.2) is proved in [2], Theorems T.7 and T.13; also (3.3) and (3.5) are proved in [3] (see Addendum, second paragraph), Lemma 3.2 and Lemma 3.7 here respectively, the last using boundedness of $V(\boldsymbol{x})$ over $b > |\boldsymbol{x}| > b-1 \geqq R_0$. (3.4) is obvious from the definition of $m(\lambda; D, q)$, and $V(\boldsymbol{x}) \geqq -g(|\boldsymbol{x}|) = q_0(\boldsymbol{x})$ almost (μ_n) everywhere over $|\boldsymbol{x}| > b \geqq R_0 + 1$ by Condition 2.B yields $\tau_p(D_b^+, V) \geqq \tau_p(D_b^+, q_0)$ in (3.1) and hence yields (3.6).

In view of (3.2)—(3.6), we see that the desired conclusion of Theorem 2.C follows if we can show $m(\lambda; D_b^+, q_0)$ bounded over $\lambda < 0$. Since Condition 2.B assures that $q_0(\boldsymbol{x}) = -g(|\boldsymbol{x}|)$ possesses continuous first partials throughout $|\boldsymbol{x}| > R_0$ and that $\lim_{\mathsf{r}\to+\infty} g(\mathsf{r}) = 0$, it follows ([3], Lemmas 3.5 and 3.6 and Addendum, second paragraph) for $b-1 \geqq R_0$, $q_0(x) = -g(|\boldsymbol{x}|)$ that over $\lambda < 0$

$$m(\lambda; D_b^+, q_0) \leqq N_b^+(\lambda; q_0) < +\infty, \tag{3.7}$$

(actually $m = N$ here). Here $N_b^+(\lambda; q_0)$ is defined (see [3] preceding Lemma 3.3 here) as the dimension of the complex linear manifold (necessarily closed since finite-dimensional) in $L_2(D_b^+)$ generated by all real-valued $u \in L_2(D_b^+)$ having continuous second partials throughout D_b^+, satisfying

$$-\nabla^2 u(\boldsymbol{x}) - g(|\boldsymbol{x}|)\, u(\boldsymbol{x}) = \lambda' u(\boldsymbol{x}) \tag{3.8}$$

throughout $\boldsymbol{x} \in D_b^+$ for some real $\lambda' \leqq \lambda$, and satisfying also the weak Neumann condition at b^+. By the weak Neumann condition at b^+ for such a u we mean for $n \geqq 2$ that u should possess a continuous extension over $\overline{D}_b^+ = \{\boldsymbol{x} \mid |\boldsymbol{x}| \geqq b\}$, that each first partial of u should be in L_2 over $b < |\boldsymbol{x}| < b+1$, and that

$$0 = \lim_{\mathsf{r}\to b^+} \int_{|\boldsymbol{x}|=\mathsf{r}} w(\boldsymbol{x})\, \mathsf{r}^{-1} \left(\sum_{j=1}^{n} x_j \frac{\partial}{\partial x_j} u(\boldsymbol{x}) \right) d\sigma_n(\boldsymbol{x}) \tag{3.9}$$

for every real-valued function w which is continuous over $b \leqq |\boldsymbol{x}| \leqq b+1$ and possesses continuous first partials throughout $b < |\boldsymbol{x}| < b+1$ which are in L_2 there. In (3.9), σ_n denotes $n-1$-dimensional "area" measure. For $n=1$ the weak Neumann condition at b^+ means simply, with $D_b^+ = (-\infty, -b) \cup (b, +\infty)$,

$$\lim_{x\to b^+} u'(x) = 0 = \lim_{x\to(-b)^-} u'(x). \tag{3.10}$$

Finally we estimate $N_b^+(\lambda; q_0)$ in (3.7) by the usual separation of variables in (3.8) (see [3], Lemma 4.1, plus Addendum). Here for integer $\nu \geqq 0$ we define

$$ {}_nf_\nu(\mathsf{r}) = g(\mathsf{r}) - (4\mathsf{r}^2)^{-1}[(n-2)^2 - 1 + 4\nu(\nu+n-2)] \tag{3.11}$$

over real $\mathsf{r} \geqq R_0$ for dimension $n \geqq 2$, and

$$ {}_1f_\nu(\mathsf{r}) = g(\mathsf{r}) \tag{3.12}$$

over such r for dimension $n=1$. We also define

$$\begin{aligned} h_n(\nu) &= \frac{(2\nu+n-2)\,\underline{|\nu+n-3}}{\underline{|\nu}\,\underline{|n-2}} \quad \text{if} \quad n \geqq 3, \quad \nu \geqq 0, \\ h_2(\nu) &= 2 \quad \text{if} \quad \nu \geqq 1, \qquad h_2(0) = 1, \\ h_1(\nu) &= 0 \quad \text{if} \quad \nu \geqq 1, \qquad h_1(0) = 2. \end{aligned} \tag{3.13}$$

Define ${}_1N_b^+(\lambda;\nu,g)$ over real $\lambda<0$ and integer $\nu\geqq 0$ as the dimension of the closed complex linear manifold in $L_2(b,+\infty)$ generated by all real-valued $w\in L_2(b,+\infty)$ which are solutions of

$$-w''(\mathrm{r})-{}_nf_\nu(\mathrm{r})\,w(\mathrm{r})=\lambda'\,w(\mathrm{r}) \tag{3.14}$$

over $(b,+\infty)$ for some real $\lambda'\leqq\lambda$, the λ' depending on w, and which also have a finite limit $w'(b^+)$ such that

$$0=w'(b^+)-(n-1)\,(2b)^{-1}\,w(b^+)\,. \tag{3.15}$$

With the definitions (3.8)—(3.15) and with $q_0(\boldsymbol{x})=-g(|\boldsymbol{x}|)$, our separation of variables statement is that, for $b\geqq R_0+1$ and g as in Condition 2.B, over $\lambda<0$

$$N_b^+(\lambda;q_0)=\sum_{\nu=0}^{\infty}h_n(\nu)\,{}_1N_b^+(\lambda;\nu,g)\,. \tag{3.16}$$

We indicate as follows the proof of (3.16) cited above (the hypotheses upon $q_0=-g$ are different there than here, but in both cases they imply that the real-valued function g is continuous over $[b,+\infty)$ and in $L_1(b,+\infty)$ and has $\left(\liminf_{\mathrm{r}\to+\infty}\{-g(\mathrm{r})\}\right)\geqq 0$, which suffice). For $n\geqq 2$ let $\{Y_{\nu,m}\}$, with $1\leqq m\leqq h_n(\nu)$, denote the $L_2(S,\sigma_n)$ orthonormal set of $h_n(\nu)$ real spherical harmonics in E_n of degree ν, where $S=\{\boldsymbol{\vartheta}\in E_n\,|\,|\boldsymbol{\vartheta}|=1\}$ is the E_n unit sphere. Letting $\lambda_{\nu,p}<\lambda_{\nu,p+1}<0$ denote the $p\geqq 1$ indexed negative λ' eigenvalues of (3.14) under (3.15) (each of multiplicity 1), and $w_{\nu,p}\in L_2(b,+\infty)$ the corresponding normalized real eigenfunction, we trivially compute that

$$u_{\nu,m,p}(\boldsymbol{x})=\left[w_{\nu,p}(|\boldsymbol{x}|)\;Y_{\nu,m}(|\boldsymbol{x}|^{-1}\boldsymbol{x})\right]|\boldsymbol{x}|^{-\frac{1}{2}(n-1)} \tag{3.17}$$

defines $u_{\nu,m,p}\in L_2(D_b^+)$ satisfying (3.8) with $\lambda'=\lambda_{\nu,p}$ and satisfying the weak Neumann condition at b^+ from (3.15). Since the $Y_{\nu,m}$ are orthogonal in $L_2(S,\sigma_n)$ for distinct ν as well as m, and since (3.14) and (3.15) make the $w_{\nu,p}$ orthonormal in $L_2(b,+\infty)$ for each ν, the (3.17) functions are orthonormal in $L_2(D_b^+)$ for m,ν,p distinct. Thus the left side of (3.16) is $\geqq$ the right side for $n\geqq 2$. Conversely, letting u be an eigenfunction satisfying (3.8)—(3.10) for λ', we see that

$$v_{\nu,m}(\mathrm{r})=\mathrm{r}^{\frac{1}{2}(n-1)}\int_S u(\mathrm{r}\,\boldsymbol{\vartheta})\;Y_{\nu,m}(\boldsymbol{\vartheta})\;d\sigma_n(\boldsymbol{\vartheta}) \tag{3.18}$$

satisfies (3.14) and (3.15) trivially, and thus $\lambda'=\lambda_{\nu,p}$ and $v_{\nu,m}=c\,w_{\nu,p}$ with $c=(v_{\nu,m},w_{\nu,p})=(u,u_{\nu,m,p})$ of (3.17). Thus, $\{Y_{\nu,m}\}$ being complete orthonormal in $L_2(S,\sigma_n)$, by the Fubini and monotone convergence theorems we have

$$\begin{aligned}\int_{D_b^+}|u|^2\,d\mu_n&=\int_b^\infty \mathrm{r}^{n-1}\left\{\int_S|u(\mathrm{r}\,\boldsymbol{\vartheta})|^2\,d\sigma_n(\boldsymbol{\vartheta})\right\}d\,\mathrm{r}\\&=\sum_{\nu=0}^{\infty}\sum_{m=1}^{h_n(\nu)}\int_b^\infty|v_{\nu,m}(\mathrm{r})|^2\,d\,\mathrm{r}\\&=\sum_{\nu\in A(\lambda')}\sum_{m=1}^{h_n(\nu)}|(u,u_{\nu,m,p})|^2,\end{aligned} \tag{3.19}$$

where $A(\lambda')=\{\text{integer } \nu\geqq 0 \mid \lambda_{\nu,p}=\lambda' \text{ for some } p\}$ and where in the last line of (3.19) $p=p_\nu$ with p_ν for each $\nu\in A(\lambda')$, the unique p having $\lambda_{\nu,p}=\lambda'$. Thus (3.19) shows that such a u is contained in the closed linear manifold in $L_2(D_b^+)$ generated by the functions (3.17), proving equality in (3.16) for $n\geqq 2$. For $n=1$, (3.16) becomes $N_b^+(\lambda; q_0)=2\,{}_1N_b^+(\lambda; 0, g)$, which is obvious from $D_b^+=(-\infty, -b)\cup(b, +\infty)$ with $b>0$.

Having established (3.16) for $q_0(\boldsymbol{x})=-g(|\boldsymbol{x}|)$, we use it with (3.7) to obtain the boundedness of $m(\lambda; D_b^+, q_0)$ over $\lambda<0$, which as noted after (3.6) suffices to prove our main Theorem 2.C. Here from (2.2) we see in (3.11) that there exists some finite integer $\nu_1\geqq 0$ in case $n\geqq 2$ such that ${}_nf_\nu(\mathrm{r})\leqq 0$ over $\mathrm{r}\geqq R_0$ and $\nu>\nu_1$. Thus $q(\mathrm{r})=-{}_nf_\nu(\mathrm{r})\geqq 0$ there shows by application of Lemma 2.E to (3.14) and (3.15) with $k=(n-1)(2b)^{-1}$ that ${}_1N_b^+(\lambda; \nu, g)=0$ over $\lambda<0$, $b\geqq R_0+1$, $\nu>\nu_1$. Hence in case $n\geqq 2$, (3.16) becomes

$$N_b^+(\lambda; q_0)=\sum_{\nu=0}^{\nu_1} h_n(\nu)\,{}_1N_b^+(\lambda; \nu, g); \tag{3.20}$$

in case $n=1$, (3.20) with $\nu_1=0$ is equivalent to (3.16), so (3.20) holds throughout $\lambda<0$ and $b\geqq R_0+1$ with fixed $\nu_1\geqq 0$ for any dimension n.

Next define ${}_1N_b(\lambda; g)$ exactly like ${}_1N_b^+(\lambda; \nu, g)$ in (3.11)–(3.15), except that in (3.14) we substitute $g(\mathrm{r})$ for the ${}_nf_\nu(\mathrm{r})$ of (3.11) and (3.12). Similarly to (3.1), we also define for $b\geqq R_0+1$, $k_b=(n-1)(2b)^{-1}$ as in (3.15), $q=-g$ or $-{}_nf_\nu$

$$\tau_p(b, q)=\sup_S\left[\inf_u\left\{k_b|u(b^+)|^2+\int_b^\infty\left(|u'(x)|^2+q(x)\,|u(x)|^2\right)dx\right\}\right] \tag{3.21}$$

with infimums and supremums taken as in (3.1) for $D=(b, +\infty)$; $m_b(\lambda; g)$ is the number of integers $p\geqq 1$ such that $\tau_p(b, -g)\leqq\lambda$, and $m_b(\lambda; {}_nf_\nu)$ is the same for $\tau_p(b, -{}_nf_\nu)\leqq\lambda$. Now it is well known ([3], equality (3.27) noted in the Addendum last paragraph with here $k_b=\gamma$ there) that $m_b(\lambda; g)={}_1N_b(\lambda; g)$ and $m_b(\lambda; {}_nf_\nu)={}_1N_b^+(\lambda; \nu, g)$ over $\lambda<0$ and $b\geqq R_0+1$. Since (3.11) and (3.12) make $-{}_nf_\nu(\mathrm{r})\geqq -g(\mathrm{r})$ over $\mathrm{r}\geqq R_0$ except when $n=2$ and $\nu=0$, clearly $m_b(\lambda; {}_nf_\nu)\leqq m_b(\lambda; g)$, and thus

$${}_1N_b^+(\lambda; \nu, g)\leqq {}_1N_b(\lambda; g) \tag{3.22}$$

over $\lambda<0$, $b\geqq R_0+1$, integer $\nu\geqq 0$ except for $\nu=0$ when $n=2$.

Now

$$\mathrm{r}\int_{\mathrm{r}}^\infty s^{-2}(\ln s)^{-2}\,ds\leqq \mathrm{r}\int_{\mathrm{r}}^\infty s^{-2}\,ds=1 \quad \text{for} \quad r\geqq e$$

shows (2.3)′ implies (2.3), and thus (2.3) always may be assumed over $\mathrm{r}\in[R_1, +\infty)$ for some real $R_1\geqq R_0+1$. Thus g satisfies (2.6) on $[R_1, +\infty)$, and Lemma 2.D yields ${}_1N_R(\lambda; g)\leqq 1$ over $\lambda<0$ and $R\geqq R_2$ for some real $R_2\geqq R_1$. Taking $b\geqq R_2$ in (3.22) and (3.20) thus gives our desired $N_b^+(\lambda; q_0)\leqq N_0<+\infty$ over $\lambda<0$, providing we can also estimate ${}_1N_b^+(\lambda; 0, g)$ when $n=2$.

To handle the exceptional case in which $\nu=0$, $n=2$, note that (2.3)′ is assumed and ${}_2f_0(\mathrm{r})=\frac{1}{4}\mathrm{r}^{-2}+g(\mathrm{r})$ in (3.11). Hence (2.6)′ is satisfied by ${}_2f_0$, and Lemma 2.D yields ${}_1N_R^+(\lambda; 0, g)\leqq 1$ over $\lambda<0$ and $R\geqq R_3$ for some real $R_3\geqq R_0+1$. Taking $b=(\max R_2, R_3)$, when $n=2$, thus finishes the proof of Theorem 2.C.

4. Discussion of Lemmas

To prove Lemma 2.E, assume, contrary to the assertion there, the existence of a real-valued solution u of (2.8) satisfying (2.9) and $0 < \int_R^\infty [u(\mathrm{r})]^2 d\,\mathrm{r} < +\infty$. Then (2.8) with q bounded also makes $u''(\mathrm{r}) \in L_2(R, \infty)$; thus it is well known ([3], equation (3.5) plus references) also that $u'(\mathrm{r}) \in L_2(R, \infty)$ and that

$$\int_R^\infty \{u(\mathrm{r})\, u''(\mathrm{r}) + [u'(\mathrm{r})]^2\}\, d\,\mathrm{r} = -\, u(R^+)\, u'(R^+) = -\, k\, [u(R^+)]^2, \tag{4.1}$$

using (2.9) also. Hence

$$\begin{aligned} 0 \geqq -\, k\, [u(R^+)]^2 &= \int_R^\infty \{[u'(\mathrm{r})]^2 + (q(\mathrm{r}) - \lambda)\, [u(\mathrm{r})]^2\}\, d\,\mathrm{r} \\ &\geqq \int_R^\infty \{q(\mathrm{r}) - \lambda\}\, [u(\mathrm{r})]^2\, d\,\mathrm{r} \geqq \{[\inf_{\mathrm{r} \geqq R} q(\mathrm{r})] - \lambda\} \int_R^\infty [u(\mathrm{r})]^2\, d\,\mathrm{r} > 0, \end{aligned} \tag{4.2}$$

an obvious contradiction. Thus Lemma 2.E follows.

Turning now to Lemma 2.D, it is well known ([4], we are indebted to Professor HARTMAN for pointing this out to us) that for f as assumed here the conclusion of Lemma 2.D is implied by the equation

$$-\, u''(\mathrm{r}) - f(\mathrm{r})\, u(\mathrm{r}) = 0 \tag{4.3}$$

being non-oscillatory at $+\infty$, by which we mean that there exists some real $R \geqq R_1$ such that every not identically vanishing real-valued solution u of (4.3) over $(R, +\infty)$ possesses at most a single zero in $(R, +\infty)$. But (2.6) is one of HILLE's sufficient conditions ([5], Theorem 7, Corollary 1 or Theorem 12, case $p = 0$) that (4.3) be non-oscillatory at $+\infty$. Likewise, since $\mathrm{r} \int_\mathrm{r}^\infty \frac{1}{4} s^{-2} d\,s = \frac{1}{4}$, (2.6)′ implies

$$\mathrm{r} \int_\mathrm{r}^\infty f(s)\, d\,s \leqq \tfrac{1}{4}\, \mathrm{r} \int_\mathrm{r}^\infty \{s^{-2} + (s \ln s)^{-2}\}\, d\,s \tag{4.4}$$

over $\mathrm{r} \geqq R_1$, which is another one ([5], Theorem 12, case $p = 1$) of HILLE's sufficient conditions that (4.3) be non-oscillatory at $+\infty$. Hence Lemma 2.D has been proved.

In closing we remark that other non-oscillation conditions have been given for (4.3), but most seem to imply those of HILLE quoted above; naturally each such condition leads to a modification of (2.3) or (2.3)′ for which our Theorem 2.C follows. If (2.6)′ is dropped and (2.6) is strengthened to $\int_{R_0}^\infty \mathrm{r}\, f(\mathrm{r})\, d\,\mathrm{r} < +\infty$, it is fairly easy to give a direct proof (not appealing to non-oscillation of (4.3)) of Lemma 2.D; for from the well known asymptotics at ∞ of solutions of (2.4), such solutions in $L_2(R, +\infty)$ are $O(e^{-\omega \mathrm{r}})$ and satisfy

$$u(\mathrm{r}) = e^{-\omega \mathrm{r}} - \int_\mathrm{r}^\infty \left\{\frac{\sinh[\omega(s - r)]}{\omega}\right\} f(s)\, u(s)\, d\,s, \tag{4.5}$$

from which the violation of (2.5) can be deduced, at least for those $\vartheta \in [0, \pi/2)$ given by (3.15).

References

[1] Brownell, F. H.: Spectrum of the static potential Schrödinger equation over E_n. Annals Math. **54**, 554—594 (1951).

[2] Brownell, F. H.: A note on Kato's uniqueness criterion for Schrödinger operator self-adjoint extensions. Pacific J. Math. **9**, 953—973 (1959).

[3] Brownell, F. H., & C. W. Clark: Asymptotic distribution of the eigenvalues of the lower part of the Schrödinger operator spectrum. J. Math. Mech. **10**, 31—70 (1961); Addendum, May, 1961.

[4] Hartman, P.: Differential equations with non-oscillatory eigenfunctions. Duke Math. J. **15**, 697—709 (1948).

[5] Hille, E.: Non-oscillation theorems. Trans. Amer. Math. Soc. **64**, 234—252 (1948).

[6] Ikebe, T.: Eigenfunction expansions for the Schrödinger operator and scattering theory. Arch. Rational Mech. Anal. **5**, 1—34 (1960).

[7] Kato, T.: Growth properties of solutions of the reduced wave equation with a variable coefficient. Comm. Pure Appl. Math. **12**, 403—425 (1959).

Department of Mathematics
University of Washington
Seattle, Washington

(Received March 27, 1961)

Difference Methods for Two-Dimensional Mixed Problems for Hyperbolic First Order Systems

VIDAR THOMÉE

Communicated by LARS HÖRMANDER

This paper is concerned with a boundary problem of mixed type for a first-order hyperbolic system of partial differential equations in two independent variables.

In two previous papers [*10*], [*11*], the author has treated the linear problem by L^2-methods. The existence proof, which is carried out in [*11*] by use of functional-analytic tools, is based on *a priori* estimates derived in [*10*]. However, it is not suited for numerical solution of the boundary-value problem. In this paper, an alternative existence proof is given by means of an explicit difference scheme. This method has the advantages of being applicable for numerical purposes and of yielding in a simple way the regularity properties of the solution. Moreover, it is suited for the diagonable quasi-linear case.

The linear problem has been treated previously by other methods by CAMPBELL & ROBINSON [*1*] and in more than two dimensions by DUFF [*4*]. PEYSER [*8*] has considered the closely connected case of a hyperbolic equation of arbitrary order in two dimensions.

Difference methods have been much used in the theory of partial differential equations (see *e.g.* RICHTMYER [*9*], with references). The success of a difference scheme is connected with the possibility of deducing *a priori* estimates for the solutions, or with the "stability" of the difference scheme. Such estimates are also the central tools in this paper. LAX & KELLER [*6*] have considered an alternative difference scheme for a similar boundary problem and proved stability under the assumption of the corresponding existence theorem. In a recent paper, KREISS [*5*] has studied the stability problem for a class of equations containing the one treated here.

From the extensive literature on the Cauchy problem for the same system the papers by COURANT, ISAACSON & REES [*2*], and by COURANT & LAX [*3*] should be mentioned. In [*2*] the same difference approximations are used as here and also are suggested for study of the stability in the case of the mixed problem. In [*3*] a method of existence proof by means of integral equations which is used for the Cauchy problem is suggested also for the mixed problem.

1. The differential and boundary operators

Let $\mathscr{V}$ be the set of points $\{(x, t) \mid x \geqq 0,\ t \geqq 0,\ x + t/\lambda \leqq 1\}$ in Euclidean two-dimensional space. Here λ is a positive constant to be specified later. Let $\mathscr{S}^{t_0}$ and $\mathscr{S}^-$ be the sets $\mathscr{V} \cap \{t = t_0\}$ and $\mathscr{V} \cap \{x = 0\}$.

Let $\mathfrak{F}(A, B)$ be the set of mappings from A into B, and let $\mathfrak{L}(A, B)$ and $\mathscr{C}^q(A, B)$ be the subsets of linear and of continuous and q-times continuously differentiable mappings. (In case $q = 0$ we usually omit this index.) Let C be the set of complex numbers and C_μ the set of vectors $v = (v_1, \ldots, v_\mu)$ (in the sequal often considered as column vectors) with components in C. Let $C_{\mu_1 \times \mu_2} = \mathfrak{L}(C_{\mu_1}, C_{\mu_2})$, *i.e.* the set of complex-valued matrices with μ_1 columns and μ_2 rows. For $\mathsf{M} = \mu$ or $\mu_1 \times \mu_2$ and $W = \mathscr{V}, \mathscr{S}^0, \mathscr{S}^-$ we put $\mathscr{C}^q_{\mathsf{M}}(W) = \mathscr{C}^q(W, C_{\mathsf{M}})$. (We often omit M and W and simply write $\mathscr{C}^q$.)

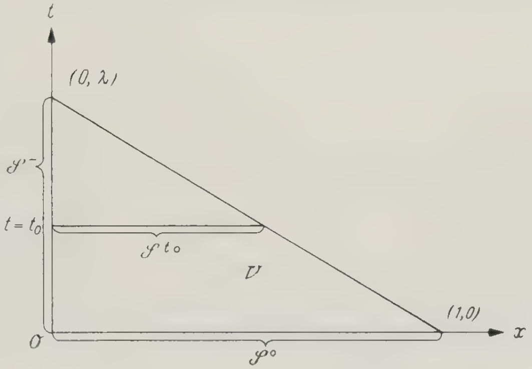

Fig. 1

Assume that $\mathscr{D}$ is a real diagonal matrix $\in \mathscr{C}^0_{m \times m}(\mathscr{V})$ and that $\mathscr{G} \in \mathscr{C}^0_{m \times m}(\mathscr{V})$. The equation

$$(1.1) \quad \mathscr{L} u = \frac{\partial u}{\partial t} - \mathscr{D} \frac{\partial u}{\partial x} + \mathscr{G} u$$

then defines a linear hyperbolic operator $\mathscr{L} \in \mathfrak{L}(\mathscr{C}^1_m(\mathscr{V}), \mathscr{C}^0_m(\mathscr{V}))$. We suppose that $\mathscr{D}$ has the form

$$\mathscr{D} = \begin{pmatrix} \mathscr{D}^+ & 0 \\ 0 & \mathscr{D}^- \end{pmatrix}$$

where $\mathscr{D}^+$ and $\mathscr{D}^-$ are diagonal matrices in $\mathscr{C}^0_{m^+ \times m^+}(\mathscr{V})$ and $\mathscr{C}^0_{m^- \times m^-}(\mathscr{V})$, which are positive and negative definite in $\mathscr{V}$, respectively, and the zeros represent blocks with only zeros. Correspondingly we can write $u = (u^+, u^-)$ where $u^+ \in \mathscr{C}^1_{m^+}(\mathscr{V})$, $u^- \in \mathscr{C}^1_{m^-}(\mathscr{V})$, and partition $\mathscr{G}$ into $\mathscr{G}^+ \in \mathscr{C}^0_{m \times m^+}(\mathscr{V})$ and $\mathscr{G}^- \in \mathscr{C}^0_{m \times m^-}(\mathscr{V})$. With a similar partitioning of $\mathscr{L}$, (1.1) can be written

$$\mathscr{L}^+ u = \frac{\partial u^+}{\partial t} - \mathscr{D}^+ \frac{\partial u^+}{\partial x} + \mathscr{G}^+ u,$$

$$\mathscr{L}^- u = \frac{\partial u^-}{\partial t} - \mathscr{D}^- \frac{\partial u^-}{\partial x} + \mathscr{G}^- u.$$

Let ϱ^0 and ϱ^- be the restriction operators from $\mathscr{V}$ to $\mathscr{S}^0$ and $\mathscr{S}^-$, respectively, so that for $u \in \mathscr{C}^0_{\mathsf{M}}(\mathscr{V})$ ($\mathsf{M} = \mu$ or $\mu_1 \times \mu_2$),

$$\varrho^0 u(x) = u(x, 0), \qquad \varrho^- u(t) = u(0, t).$$

On $\mathscr{S}^0$ we consider the initial-value operator

$$\ell^0 u = \varrho^0 u, \qquad \ell^0 \in \mathfrak{L}(\mathscr{C}^0_m(\mathscr{V}), \mathscr{C}^0_m(\mathscr{S}^0)),$$

and on $\mathscr{S}^-$ the boundary operator

$$\ell^- u = \mathscr{B} \varrho^- u, \qquad \ell^- \in \mathfrak{L}(\mathscr{C}^0_m(\mathscr{V}), \mathscr{C}^0_{m^-}(\mathscr{S}^-)),$$

where $\mathscr{B}$ has the partitioning $\mathscr{B} = (\mathscr{B}^+, I^-)$ with $\mathscr{B}^+ \in \mathscr{C}^0_{m^+ \times m^-}(\mathscr{S}^-)$ and I^- is the identity matrix of the type $m^- \times m^-$. With these notations the problem to be considered can be formulated: Given $\mathscr{F}, f^0, f^-$. Find $u \in \mathscr{C}_m^1(\mathscr{V})$ such that

$$\begin{aligned} \mathscr{L} u &= \mathscr{F}, \\ \ell^0 u &= f^0, \qquad (1.2) \\ \ell^- u &= f^-. \end{aligned}$$

2. The difference operators

We shall approximate $\mathscr{L}$ by means of a difference operator. We introduce for arbitrary n the grid

$$V = \{(\xi h, \tau k) \mid \tau = 0, \dots, n;\ \xi = 0, \dots, n - \tau;\ h = 1/n;\ k = \lambda/n\}$$

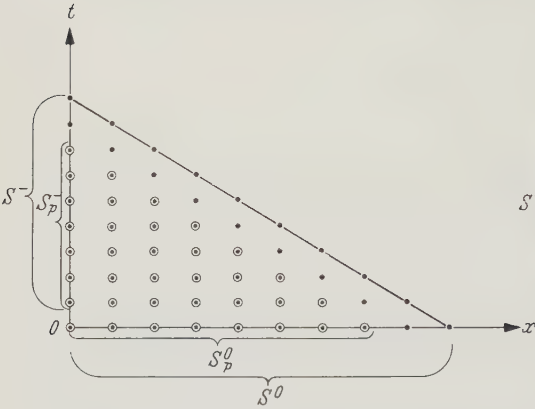

Fig. 2. • = points of V; ○ = points of V_p ($n=10$; $p=2$)

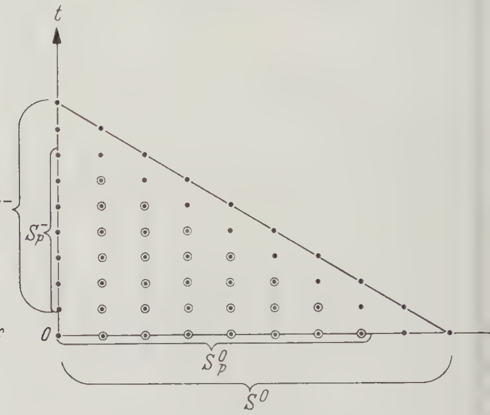

Fig. 3. • = points of V; ○ = points of V_p^- ($n=10$; $p=2$)

and its intersections with $\mathscr{S}^{\tau k}$ and $\mathscr{S}^-$,

$$\begin{aligned} S^\tau &= \{(\xi h, \tau k) \mid \xi = 0, \dots, n - \tau\}, \qquad 0 \leqq \tau \leqq n, \\ S^- &= \{(0, \tau k) \mid \tau = 1, \dots, n\}. \end{aligned}$$

We shall need also

$$\begin{aligned} V^- &= \{(\xi h, \tau k) \mid \tau = 0, \dots, n - 1;\ \xi = 1, \dots, n - \tau\}, \\ V_p &= \{(\xi h, \tau k) \mid \tau = 0, \dots, n - p;\ \xi = 0, \dots, n - p - \tau\}, \\ V_p^- &= V_p \cap V^-, \qquad S_p^- = V_p \cap S^-, \qquad S_p^\tau = V_p \cap S^\tau. \end{aligned}$$

For $\mathsf{M} = \mu$ or $\mu_1 \times \mu_2$ and $W = V$, S^0 *etc.* we put $\mathscr{C}_{\mathsf{M}}(W) = \mathfrak{F}(W, C_{\mathsf{M}})$. (We often omit the index M.) For $v \in \mathscr{C}_{\mathsf{M}}(V)$ we now define (for brevity we write $v_{\xi,\tau}$ instead of $v(\xi h, \tau k)$, *etc.*)

$$\begin{aligned} \Delta_\tau v_{\xi,\tau} &= k^{-1}(v_{\xi,\tau+1} - v_{\xi,\tau}), \\ \Delta_\xi v_{\xi,\tau} = \Delta_\xi^+ v_{\xi,\tau} &= h^{-1}(v_{\xi+1,\tau} - v_{\xi,\tau}), \\ \Delta_\xi^- v_{\xi,\tau} &= \Delta_\xi^+ v_{\xi-1,\tau}. \end{aligned}$$

Here Δ_τ and $\Delta_\xi \in \mathfrak{L}(\mathscr{C}(V), \mathscr{C}(V_1))$ and $\Delta_\xi^- \in \mathfrak{L}(\mathscr{C}(V), \mathscr{C}(V_1^-))$. The corresponding definitions of Δ_ξ and Δ_τ for $v \in \mathscr{C}(S^0)$ and $\mathscr{C}(S^-)$ are obvious.

Let R_p, R_p^-, r^0, and r^- (p omitted when $=0$) be the restriction operators from $\mathscr{V}$, $\mathscr{V}$, $\mathscr{S}^0$, and $\mathscr{S}^-$ to V_p, V_p^-, S^0, and S^-, respectively, and σ^0 and σ^- the restriction operators from V to S^0 and S^- so that for instance if $A \in \mathscr{C}(\mathscr{V})$, then $R_p A \in \mathscr{C}(V_p)$ and $R_p A_{\xi,\tau} = A(\xi h, \tau k)$ if $(\xi h, \tau k) \in V_p$.

We shall now define the difference approximations of $\mathscr{L}^+$ and $\mathscr{L}^-$. We put $D^+ = R_1 \mathscr{D}^+$, $D^- = R_1^- \mathscr{D}^-$, $G^+ = R_1 \mathscr{G}^+$, and $G^- = R_1^- \mathscr{G}^-$ and define for $u \in \mathscr{C}_m(V)$ $\big(u = (u^+, u^-)$, $u^+ \in \mathscr{C}_{m^+}(V)$, $u^- \in \mathscr{C}_{m^-}(V)\big)$:

$$
\begin{aligned}
L^+ u &= \Delta_\tau u^+ - D^+ \Delta_\xi^+ u^+ + G^+ u, \\
L^- u &= \Delta_\tau u^- - D^- \Delta_\xi^- u^- + G^- u.
\end{aligned} \tag{2.1}
$$

Here $L^+ \in \mathfrak{L}\big(\mathscr{C}_m(V), \mathscr{C}_{m^+}(V_1)\big)$ and $L^- \in \mathfrak{L}\big(\mathscr{C}_m(V), \mathscr{C}_{m^-}(V_1^-)\big)$. We also introduce operators corresponding to ℓ^0 and ℓ^-. We put for $u \in \mathscr{C}_m(V)$ with $B = r^- \mathscr{B}$

$$
l^0 u = \sigma^0 u, \qquad l^- u = B \sigma^- u.
$$

Here $l^0 \in \mathfrak{L}\big(\mathscr{C}_m(V), \mathscr{C}_m(S^0)\big)$ and $l^- \in \mathfrak{L}\big(\mathscr{C}_m(V), \mathscr{C}_{m^-}(S^-)\big)$.

In addition to the boundary problem (1.2) we shall consider the discrete boundary problem: Given $\mathscr{F} = (\mathscr{F}^+, \mathscr{F}^-) \in \mathscr{C}_m(\mathscr{V})$, $\mathscr{f}^0 \in \mathscr{C}_m(\mathscr{S}^0)$, $\mathscr{f}^- \in \mathscr{C}_{m^-}(\mathscr{S}^-)$, find $u \in \mathscr{C}_m(V)$ such that

$$
\begin{aligned}
L^+ u &= F^+, && \text{where} \quad F^+ = R_1 \mathscr{F}^+, \\
L^- u &= F^-, && \text{where} \quad F^- = R_1^- \mathscr{F}^-, \\
l^0 u &= f^0, && \text{where} \quad f^0 = r^0 \mathscr{f}^0, \\
l^- u &= f^-, && \text{where} \quad f^- = r^- \mathscr{f}^-.
\end{aligned} \tag{2.2}
$$

We have the following theorem:

Theorem 2.1. *There is one and only one solution* $u \in \mathscr{C}_m(V)$ *of the discrete boundary problem* (2.2).

Proof. Theorem 2.1 follows from the following explicit scheme, which is easily obtained from the definitions:

$$
u_{\xi,0} = l^0 u_\xi, \qquad \xi = 0, 1, \dots, n; \tag{2.3}
$$

$$
\begin{gathered}
u^+_{\xi,\tau+1} = \lambda D^+_{\xi,\tau} u^+_{\xi+1,\tau} + (I^+ - \lambda D^+_{\xi,\tau}) u^+_{\xi,\tau} - k G^+_{\xi,\tau} u_{\xi,\tau} + k L^+ u_{\xi,\tau}, \\
\xi = 0, 1, \dots, n - \tau - 1;
\end{gathered} \tag{2.4}
$$

$$
\begin{gathered}
u^-_{\xi,\tau+1} = (I^- + \lambda D^-_{\xi,\tau}) u^-_{\xi,\tau} - \lambda D^-_{\xi,\tau} u^-_{\xi-1,\tau} - k G^-_{\xi,\tau} u_{\xi,\tau} + k L^- u_{\xi,\tau}, \\
\xi = 1, \dots, n - \tau - 1;
\end{gathered} \tag{2.5}
$$

$$
u^-_{0,\tau+1} = l^- u_{\tau+1} - B^+_{\tau+1} u^+_{0,\tau+1}. \tag{2.6}
$$

3. Norms

We define for a vector $\varphi = (\varphi_1, \dots, \varphi_\mu) \in \mathscr{C}_\mu(W)$ or a matrix $\varphi = (\varphi_{ij}) \in \mathscr{C}_{\mu_1 \times \mu_2}(W)$ $(W = \mathscr{V}$, V_p, S^0, $0 = \mathscr{O} =$ the point $(0, 0)$, *etc.*$)$

$$
|\varphi|_W = \sup_{j, W} |\varphi_j|; \qquad |\varphi|_W = \sup_{i, j, W} |\varphi_{ij}|.
$$

Let α_x and α_t be non-negative integers, $\alpha=(\alpha_x, \alpha_t)$, $|\alpha|=\alpha_x+\alpha_t$ and $\partial^\alpha=\left(\frac{\partial}{\partial x}\right)^{\alpha_x}\left(\frac{\partial}{\partial t}\right)^{\alpha_t}$, $\Delta^\alpha=\Delta_\xi^{\alpha_x}\Delta_\tau^{\alpha_t}$. We put

$$(3.1)\qquad |\varphi|_{W,q}=\sup_{|\alpha|\leqq q}|\partial^\alpha\varphi|_W; \qquad W=\mathscr{V}, \mathscr{S}^0, \mathscr{S}^-.$$

$$(3.2)\qquad |\varphi|_{W_p,q}=\sup_{|\alpha|\leqq q}|\Delta^\alpha\varphi|_{W_{p+|\alpha|}}, \qquad W=V, V^-, S^0, S^-,$$

$$|\varphi|_{\mathscr{O},q}=\sup_{|\alpha|<q}|\partial^\alpha\varphi|_{\mathscr{C}}, \qquad |\varphi|_{0,q}=\sup_{|\alpha|\leqq q}|\Delta^\alpha\varphi|_0.$$

In the cases $W=\mathscr{S}^0(S^0)$ and $\mathscr{S}^-(S^-)$, the sup is taken only over α with $\alpha_t=0$ and $\alpha_x=0$, respectively. The derivatives in (3.1) are in these cases ordinary derivatives. (When $p=0$ in (3.2), it is usually omitted.)

We shall state some simply proved properties of these norms.

Lemma 3.1. *For each q, there is a constant C_q, independent of n, p and of $a_i\in\mathscr{C}_1(W)$, $i=1, 2$; $W=V_p, V_p^-, S_p^0, S_p^-$ such that*

$$|a_1 a_2|_{W,q}\leqq C_q|a_1|_{W,q}|a_2|_{W,q}.$$

Lemma 3.2. *If $a\in\mathscr{C}^q(W)$, $W=\mathscr{V}$, $\mathscr{V}$, $\mathscr{S}^0$, and $\mathscr{S}^-$, respectively, then*

$$|R_p a|_{V_p,q}\leqq|a|_{\mathscr{V},q};$$

$$|R_p^- a|_{V_p^-,q}\leqq|a|_{\mathscr{V},q};$$

$$|r_p^0 a|_{S_p^0,q}\leqq|a|_{\mathscr{S}^0,q};$$

$$|r_p^- a|_{S_p^-,q}\leqq|a|_{\mathscr{S}^-,q}.$$

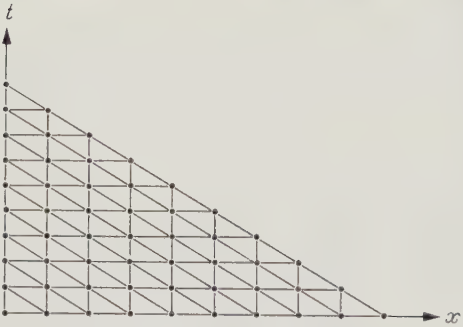

Fig. 4

For the functions $A\in\mathscr{C}_{\mathsf{M}}(V_p)$ we define an extension operator $E\in\mathfrak{L}(\mathscr{C}_{\mathsf{M}}(V_p), \mathscr{C}_{\mathsf{M}}(\mathscr{V}))$ in the following way. We first extend (if necessary) A to the whole of V by putting

$$A_{\xi,\tau}=A_{n-p-\tau,\tau} \quad \text{for} \quad \tau=0,\dots,n-p; \qquad \xi=n-p-\tau+1,\dots,n-\tau,$$

$$A_{\xi,\tau}=A_{0,n-p} \quad \text{for} \quad \tau=n-p+1,\dots,n; \qquad \xi=0,\dots,n-\tau.$$

Then we triangulate V by means of the $3(n+1)$ straight lines $x/h=j$; $t/k=j$; $x/h+t/k=j$; $j=0,\dots,n$. In each of the triangles the value of EA is determined by the values in the three edges so that $EA(\xi h, \tau k)=A_{\xi,\tau}$ and so that EA is linear in the triangle. Similarly, we define the extension operator E in the case $A^-\in\mathscr{C}_{\mathsf{M}}(V_p^-)$. In this case we extend A^- to the whole of V by putting

$$A_{\xi,\tau}^-=A_{n-p-\tau,\tau}^-, \quad \text{for} \quad \tau=0,\dots,n-p-1, \qquad \xi=n-p-\tau+1,\dots,n-\tau,$$

$$A_{0,\tau}^-=A_{1,\tau}^-, \quad \text{for} \quad \tau=0,\dots,n-p-1,$$

$$A_{\xi,\tau}^-=A_{1,n-p-1}^-, \quad \text{for} \quad \tau=n-p,\dots,n; \qquad \xi=0,\dots,n-\tau,$$

and then proceed as before. In the same way we define extension operators $E^0\in\mathfrak{L}(\mathscr{C}_{\mathsf{M}}(S_p^0), \mathscr{C}_{\mathsf{M}}(\mathscr{S}^0))$ and $E^-\in\mathfrak{L}(\mathscr{C}_{\mathsf{M}}(S_p^-), \mathscr{C}_{\mathsf{M}}(\mathscr{S}^-))$ by linear interpolation such that $E^0a^0(\xi h)=a_\xi^0$ and $E^-a^-(\tau k)=a_\tau^-$ where the definitions of a_ξ^0 and a_τ^- are extended through

$$a_\xi^0 = a_{n\ \ p}^0, \qquad \xi = n - p + 1, \ldots, n,$$
$$a_0^- = a_1^-,$$
$$a_\tau^- = a_{n-p}^-, \qquad \tau = n - p + 1, \ldots, n.$$

Some of the important properties of the extension operators are described in the following lemmas.

Lemma 3.3. *If* $A \in \mathscr{C}_{\mathrm{M}}(V_p)$, $A\ \ \in \mathscr{C}_{\mathrm{M}}(V_p^-)$, $a^0 \in \mathscr{C}_{\mathrm{M}}(S_p^0)$, *and* $a^- \in \mathscr{C}_{\mathrm{M}}(S_p^-)$, *then*

$$|EA|_{\mathscr{V}} = |A|_{V_p}; \qquad |EA\ \ |_{\mathscr{V}} = |A\ \ |_{V_p^-}; \qquad |E^0 a^0|_{\mathscr{S}^0} = |a^0|_{S_p^0}; \qquad |E^- a^-|_{\mathscr{S}^-} = |a^-|_{S_p^-}.$$

Lemma 3.4. *If* $a \in \mathscr{C}_{\mathrm{M}}^q(W)$, *then for* $n \to \infty$, *we have in* $\mathscr{C}_{\mathrm{M}}(W)$, $W = \mathscr{V}, \mathscr{V}, \mathscr{S}^0, \mathscr{S}^-$,

$$E\Delta^\alpha R_p a \to \partial^\alpha a; \qquad |\alpha| \leqq q;$$
$$E\Delta^\alpha R_p^- a \to \partial^\alpha a; \qquad |\alpha| \leqq q;$$
$$E^0\Delta^\alpha r^0 a \to \partial^\alpha a; \qquad |\alpha| \leqq q; \quad \alpha_t = 0;$$
$$E^-\Delta^\alpha r^- a \to \partial^\alpha a; \qquad |\alpha| \leqq q; \quad \alpha_x = 0.$$

(Convergence in $\mathscr{C}_{\mathrm{M}}(W)$ means uniform convergence in W.)

Lemma 3.5. *If* $a \in \mathscr{C}_1^2(\mathscr{V})$, *then* $|ERa - a|_{\mathscr{V}} \leqq \frac{1}{2}(h+k)^2 |a|_{\mathscr{V},2}$.

Lemma 3.6. *If* $a_{i,n} \in \mathscr{C}(V_p)$, $i = 1, 2$, $n = 1, 2, \ldots$ *and* $E a_{i,n} \to a_i$ *in* $\mathscr{C}(\mathscr{V})$ *when* $n \to \infty$, *then* $E(a_{1,n}\, a_{2,n}) \to a_1 a_2$ *in* $\mathscr{C}(\mathscr{V})$ *when* $n \to \infty$. *The corresponding statements hold for* V_p^-, S_p^0, *and* S_p^-.

The following two lemmas will be used in the existence proofs.

Lemma 3.7. *If for every* n, $u_n \in \mathscr{C}_{\mathrm{M}}(V)$ *and* $|u_n|_{V,1} \leqq C$, *where* C *is a constant, then a sequence* $\{n_j\}_1^\infty$ *can be found such that* $E u_{n_j}$ *converges to a function* u *in* $\mathscr{C}_{\mathrm{M}}(\mathscr{V})$ *when* $j \to \infty$.

Lemma 3.8. *If for every* n, $u_n \in \mathscr{C}_{\mathrm{M}}(V)$ *and* $|u_n|_{V,q+1} \leqq C$, *where* $q > 0$ *and* C *is a constant, then a sequence* $\{n_j\}_1^\infty$, *can be found such that* $E u_{n_j}$ *converges to a function* u *in* $\mathscr{C}_{\mathrm{M}}^q(\mathscr{V})$ *and* $E\Delta^\alpha u_{n_j}$ *converges uniformly to* $\partial^\alpha u$ *for* $|\alpha| \leqq q$ *when* $j \to \infty$.

Lemmas 3.7 and 3.8 are consequences of ARZELÀ's theorem (*cf.* PETROWSKI [7], p. 278 ff.).

4. Estimates

In Sections 4—6 we shall make the following assumption about the constant λ:

$$0 < \lambda \leqq |\mathscr{D}|_{\mathscr{V}}^{-1}.$$

The main result in this section is the following theorem.

Theorem 4.1. *There is a constant* C, *independent of* n *and* $u \in \mathscr{C}_m(V)$, *such that*

$$|u|_V \leqq C\{|L^+ u|_{V_1} + |L^- u|_{V_1^-} + |l^0 u|_{S^0} + |l^- u|_{S_8^-}\}.$$

For the proof we need Lemma 4.2, which is proved by aid of Lemma 4.1.

Lemma 4.1. *Assume that* $a_\tau \geqq 0$, $\tau = 0, \ldots, n$, *and*

$$a_{\tau+1} \leqq a_\tau + n^{-1} C_0 (a_\tau + K), \qquad \tau = 0, 1, \ldots, n-1.$$

Then for $\delta > 0$ *and* $\tau \leqq \delta n$,

$$a_\tau \leqq e^{C_0 \delta} a_0 + (e^{C_0 \delta} - 1) K .$$

Proof. We have

$$a_{\tau+1} + K \leqq (1 + C_0 n^{-1}) (a_\tau + K) \leqq e^{C_0/n} (a_\tau + K) ,$$

and hence for $\tau \leqq \delta n$,

$$a_\tau + K \leqq e^{C_0 \delta} (a_0 + K) .$$

Lemma 4.2. *Assume that* $a_\tau^+ \geqq 0$, $a_\tau^- \geqq 0$, $a_\tau = \max(a_\tau^+, a_\tau^-)$, $\tau = 0, \ldots, n$ *and that*

$$\begin{aligned} &a_{\tau+1}^+ \leqq a_\tau^+ + n^{-1} C_1 (a_\tau + K) , && \tau = 0, 1, \ldots, n-1 , \\ &a_{\tau+1}^- \leqq \max\{a_\tau^- + n^{-1} C_2 (a_\tau + K) ; C_3 (a_{\tau+1}^+ + K)\} , && \tau = 0, 1, \ldots, n-1 . \end{aligned} \tag{4.1}$$

Then there exists a constant C_4, *independent of* n, *such that*

$$a_\tau \leqq C_4 (a_0 + K) , \qquad \tau = 0, 1, \ldots, n . \tag{4.2}$$

Proof. Without restricting generality we may suppose that

$$C_1 \geqq C_2 / C_3 . \tag{4.3}$$

We put

$$\begin{aligned} &b_0^\pm = a_0^\pm , && \\ &b_\tau = \max(b_\tau^+, b_\tau^-) , && \tau = 0, 1, \ldots, n , \\ &b_{\tau+1}^+ = b_\tau^+ + n^{-1} C_1 (b_\tau + K) , && \tau = 0, 1, \ldots, n-1 , \\ &b_{\tau+1}^- = \max\{b_\tau^- + n^{-1} C_2 (b_\tau + K) , C_3 (b_{\tau+1}^+ + K)\} , && \tau = 0, 1, \ldots, n-1 . \end{aligned} \tag{4.4}$$

It follows at once that $a_\tau^\pm \leqq b_\tau^\pm$, $a_\tau \leqq b_\tau$. Since $b_0 = a_0$, it follows that (4.2) is proved by the following inequality:

$$b_\tau \leqq C (b_0 + K) , \qquad \tau \leqq n .$$

Suppose that for some τ in (4.4), $b_\tau^- = C_3 (b_\tau^+ + K)$. Then we also have $b_{\tau+1}^- = C_3 (b_{\tau+1}^+ + K)$, for according to (4.3) we have for this τ

$$\begin{aligned} b_\tau^- + n^{-1} C_2 (b_\tau + K) &= C_3 \{b_\tau^+ + n^{-1} (C_2/C_3) (b_\tau + K) + K\} \\ &\leqq C_3 \{b_\tau^+ + n^{-1} C_1 (b_\tau + K) + K\} = C_3 (b_{\tau+1}^+ + K) . \end{aligned}$$

It follows that there is a τ_0 such that

$$\begin{aligned} &b_{\tau+1}^- = b_\tau^- + n^{-1} C_2 (b_\tau + K) , && \tau < \tau_0 , \\ &b_{\tau+1}^- = C_3 (b_{\tau+1}^+ + K) , && \tau \geqq \tau_0 . \end{aligned}$$

The case where $b_\tau^- \neq C_3 (b_\tau^+ + K)$ for $\tau = 0, \ldots, n$ also fits in here if we agree to put $\tau_0 = n$ in this case. For $\tau < \tau_0$ we thus have

$$b_{\tau+1} \leqq b_\tau + n^{-1} C_4 (b_\tau + K) .$$

According to Lemma 4.1, since $\tau_0 \leqq n$, we have

$$b_\tau \leqq e^{\tau_0 n^{-1} C_4} (b_0 + K) \leqq e^{C_4} (b_0 + K) = C_5 (b_0 + K) , \qquad \tau \leqq \tau_0 . \tag{4.5}$$

We obtain

$$b^+_{\tau_0+1} = b^+_{\tau_0} + n^{-1} C_1 (b_{\tau_0} + K) \leqq C_6 (b_{\tau_0} + K). \tag{4.6}$$

For $\tau > \tau_0$ we have

$$b^+_{\tau+1} \leqq b^+_\tau + n^{-1} C_1 (b^+_\tau + b^-_\tau + K) \leqq b^+_\tau + n^{-1} C_1 (1 + C_3)(b^+_\tau + K),$$

and according to Lemma 4.1

$$b^+_\tau \leqq C_7 (b^+_{\tau_0+1} + K);$$

that is, because of (4.6),

$$b^+_\tau \leqq C_8 (b_{\tau_0} + K), \qquad \tau > \tau_0.$$

Since

$$b^-_\tau = C_3 (b^+_\tau + K), \qquad \tau > \tau_0,$$

it follows that

$$b_\tau \leqq C_9 (b_{\tau_0} + K), \qquad \tau > \tau_0,$$

and hence according to (4.5)

$$b_\tau \leqq C (b_0 + K), \qquad \tau \leqq n,$$

which completes the proof of Lemma 4.2.

Proof of Theorem 4.1. If for $u \in \mathscr{C}_m(V)$ we put

$$a^+_\tau = |u^+|_{S^\tau}, \qquad a^-_\tau = |u^-|_{S^\tau}, \qquad a_\tau = |u|_{S^\tau},$$
$$K = |L^+ u|_{V_1} + |L^- u|_{V_1^-} + |l^- u|_{S^-},$$

then from (2.3), (2.4), and (2.5), with appropriate C_1, C_2, C_3 we obtain the inequalities (4.1), and thus the theorem follows from Lemma 4.2.

Theorem 4.1 implies the following theorem for the differential operator.

Theorem 4.2. *There is a constant* C, *independent of* $u \in \mathscr{C}^1_m(\mathscr{V})$, *such that*

$$|u|_{\mathscr{V}} \leqq C\{|\mathscr{L} u|_{\mathscr{V}} + |\ell^0 u|_{\mathscr{S}^0} + |\ell^- u|_{\mathscr{S}^-}\}. \tag{4.7}$$

Proof. We have according to Theorem 4.1

$$|R u|_V \leqq C\{|L^+ R u|_{V_1} + |L^- R u|_{V_1^-} + |l^0 R u|_{S^0} + |l^- R u|_{S^-}\}.$$

Therefore because of Lemma 3.3

$$|E R u|_{\mathscr{V}} \leqq C\{|E L^+ R u|_{\mathscr{V}} + |E L^- R u|_{\mathscr{V}} + |E^0 l^0 R u|_{\mathscr{S}^0} + |E^- l^- R u|_{\mathscr{S}^-}\},$$

and hence (4.7) follows from Lemma 3.4 and Lemma 3.6.

For later use we state the following two lemmas, which are obtained by examining the proofs of Lemma 4.2 and Theorem 4.2.

Lemma 4.3. *Assume that* $a^+_\tau \geqq 0$, $a^-_\tau \geqq 0$, $a_\tau = \max(a^+_\tau, a^-_\tau)$, $\tau = 0, \ldots, n$, *and that*

$$a^+_{\tau+1} \leqq a^+_\tau + n^{-1} C_1 (a_\tau + K_1), \qquad \tau = 0, 1, \ldots, n-1;$$
$$a^-_{\tau+1} \leqq \max\{a^-_\tau + n^{-1} C_2 (a_\tau + K_1); C_3 (a^+_{\tau+1} + K_2)\}, \qquad \tau = 0, 1, \ldots, n-1.$$

Then there exists a constant C_4, independent of n, such that for $0 \leqq \delta \leqq 1$

$$a_\tau^+ \leqq C_4(a_0 + \delta K_1 + \delta K_2),$$
$$a_\tau^- \leqq C_4(a_0 + \delta K_1 + K_2), \qquad \tau \leqq \delta n.$$

Lemma 4.4. *Let $0 \leqq \delta \leqq 1$ and $\mathscr{V}^\delta = \mathscr{V} \cap \{0 \leqq t \leqq \delta\lambda\}$, $(\mathscr{S}^-)^\delta = \mathscr{S}^- \cap \mathscr{V}^\delta$. Then there is a constant C, independent of $u \in \mathscr{C}_m^1(\mathscr{V})$, such that*

$$|u^+|_{\mathscr{V}^\delta} \leqq C\{|\ell^0 u|_{\mathscr{S}^0} + \delta|\mathscr{L} u|_{\mathscr{V}^\delta} + \delta|\ell^- u|_{(\mathscr{S}^-)^\delta}\},$$
$$|u^-|_{\mathscr{V}^\delta} \leqq C\{|\ell^0 u|_{\mathscr{S}^0} + \delta|\mathscr{L} u|_{\mathscr{V}^\delta} + |\ell^- u|_{(\mathscr{S}^-)^\delta}\}.$$

Theorem 4.3. *Assume that $\mathscr{D}, \mathscr{G}, \mathscr{B} \in \mathscr{C}^q$. Then there is a constant C, independent of n and of $u \in \mathscr{C}_m(V)$, such that*

$$|u|_{V,q} \leqq C\{|L^+ u|_{V_1,q} + |L^- u|_{V_1^-,q} + |l^0 u|_{S^0,q} + |l^- u|_{S^-,q} + |u^-|_{0,q}\}. \tag{4.8}$$

Proof. We shall prove the theorem by induction over q. For $q=0$, the statement is that of Theorem 4.1. One easily proves the following representation of the successive differences with respect to τ:

$$\begin{aligned} &\Delta_\tau^j u_{\xi,\tau+1}^+ = \lambda D_{\xi,\tau+j}^+ \Delta_\tau^j u_{\xi+1,\tau}^+ + (I^+ - \lambda D_{\xi,\tau+j}^+)\Delta_\tau^j u_{\xi,\tau}^+ + k P_{j,\xi,\tau}^+; \qquad \xi = 0, 1, \ldots \\ &\Delta_\tau^j u_{\xi,\tau+1}^- = (I^- + \lambda D_{\xi,\tau+j}^-)\Delta_\tau^j u_{\xi,\tau}^- - \lambda D_{\xi,\tau+j}^- \Delta_\tau^j u_{\xi-1,\tau}^- + k P_{j,\xi,\tau}^-; \qquad \xi = 1, 2, \ldots \\ &\Delta_\tau^j u_{0,\tau+1}^- = \Delta_\tau^j(l^- u_{\tau+1} - B_{\tau+1}^+ u_{0,\tau+1}^+). \end{aligned} \tag{4.9}$$

Here $P_{0,\xi,\tau}^\pm = L^\pm u_{\xi,t} - G_{\xi,\tau}^\pm u_{\xi,\tau}$, and for $j > 0$

$$P_{j+1,\xi,\tau}^\pm = \Delta_\tau P_{j,\xi,\tau}^\pm + \Delta_\tau D_{\xi,\tau+j}^\pm \Delta_\tau^j\{(D_{\xi,\tau}^\pm)^{-1}(\Delta_\tau u_{\xi,\tau}^\pm + G_{\xi,\tau}^\pm u_{\xi,\tau} - L^\pm u_{\xi,\tau})\}.$$

We have

$$|P_{j,\xi,\tau}^\pm| \leqq C\{|\Delta_\tau^j u_{\xi,\tau}^\pm| + |u|_{V,j-1} + |L^+ u|_{V_1,j} + |L^- u|_{V_1^-,j}\}. \tag{4.10}$$

Now assume that the theorem is proved for $q = j-1$. (4.9) and (4.10) then give

$$|\Delta_\tau^j u^+|_{S_j^{\tau+1}} \leqq |\Delta_\tau^j u^+|_{S_j^\tau} + k C_1(|\Delta_\tau^j u|_{S_j^\tau} + |L^+ u|_{V_1,j} + |L^- u|_{V_1^-,j}),$$
$$|\Delta_\tau^j u^-|_{S_j^{\tau+1}} \leqq \max\{|\Delta_\tau^j u^-|_{S_j^\tau} + k C_2(|\Delta_\tau^j u|_{S_j^\tau} + |L^+ u|_{V_1,j} + |L\ u|_{V_1^-,j});$$
$$C_3(|\Delta_\tau^j u^+|_{S_j^{\tau+1}} + |l^- u|_{S^-,j} + |L^+ u|_{V_1,j} + |L^- u|_{V_1^-,j})\},$$

and thus according to Lemma 4.2

$$|\Delta_\tau^j u|_{V_j} \leqq C\{|\Delta_\tau^j u|_{S_j^0} + |L^+ u|_{V_1,j} + |L^- u|_{V_1^-,j} + |l^- u|_{S^-,j}\}. \tag{4.11}$$

Now using (2.1), we easily obtain

$$|\Delta_\tau^j u|_{S_j^0} \leqq C\{|l^0 u|_{S^0,j} + |L^+ u|_{V_1,j-1} + |L^- u|_{V_1^-,j-1} + |u^-|_{0,j} + |u|_{V,j-1}\}. \tag{4.12}$$

Using successively the formulas (2.1) and, in the neighborhood of the line $\xi + \tau = n$, the easily proved formula

$$\begin{aligned} \Delta_\xi u_{\xi,\tau}^- = (I^- + \lambda D_{\xi+1,\tau-1}^-)\Delta_\xi u_{\xi,\tau-1}^- + \lambda \Delta_\tau u_{\xi,\tau-1}^- - \\ - \lambda G_{\xi+1,\tau-1}^- u_{\xi+1,\tau-1} + \lambda L^- u_{\xi+1,\tau-1}, \qquad \tau > 0, \end{aligned} \tag{4.13}$$

we obtain for $0 \leqq l \leqq j$

$$(4.14)\qquad |\Delta_\tau^l \Delta_\xi^{j-l} u|_{V_j} \leqq C\{|\Delta_\tau^j u|_{V_j} + |L^+ u|_{V_1, j-1} + |L^- u|_{V_1^-, j-1} + |u|_{V, j-1}\}.$$

The induction assumption (4.11), (4.12), and (4.14) now prove (4.8).

Theorem 4.3 implies the following theorem for the differential operators.

Theorem 4.4. *Assume that $\mathscr{D}, \mathscr{G}, \mathscr{B} \in \mathscr{C}^q$. Then there is a constant C, independent of $u \in \mathscr{C}_m^{q+1}(\mathscr{V})$, such that*

$$(4.15)\qquad |u|_{\mathscr{V}, q} \leqq C\{|\mathscr{L} u|_{\mathscr{V}, q} + |\ell^0 u|_{\mathscr{S}^0, q} + |\ell^- u|_{\mathscr{S}^-, q}\}.$$

Theorem 4.4 is proved in the same way as Theorem 4.2. The term $|u^-|_{\mathscr{O}, q}$ which is obtained from the term $|u|_{0,q}$ in (4.8) can be estimated by means of the terms on the right-hand side of (4.15).

5. Compatibility conditions

We say that $\mathscr{F} \in \mathscr{C}_m^{q-1}(\mathscr{V})$, $f^0 \in \mathscr{C}_m^q(\mathscr{S}^0)$, $f^- \in \mathscr{C}_{m^-}^q(\mathscr{S}^-)$ satisfy compatibility conditions of order q at the origin (for the boundary problem (1.2)) if there is a $\bar{u} \in \mathscr{C}_m^q(\mathscr{V})$ such that

$$(5.1)\qquad \partial^\alpha(\mathscr{L}\bar{u})(0,0) = \partial^\alpha \mathscr{F}(0,0), \qquad |\alpha| \leqq q-1,$$

$$(5.2)\qquad d^\beta(\ell^0 \bar{u})(0) = d^\beta f^0(0), \qquad \beta \leqq q, \quad \left(d = \frac{d}{dx}\right),$$

$$(5.3)\qquad d^\beta(\ell^- \bar{u})(0) = d^\beta f^-(0), \qquad \beta \leqq q, \quad \left(d = \frac{d}{dt}\right).$$

If for instance $\mathscr{F}$, f^0, and f^- vanish in the neighborhood of the origin, then they satisfy compatibility conditions of order q for every q at the origin. We have only to put $\bar{u} = 0$ in $\mathscr{V}$. In order to define the compatibility conditions it would only be necessary to require regularity of $\mathscr{F}$, f^0, and f^- at the origin. However, this fact will not be emphasized since the global regularity will be needed in the existence proofs.

We obtain at once

Theorem 5.1. *A necessary condition for the solvability of the boundary problem (1.2) in $\mathscr{C}_m^q(\mathscr{V})$ is that $\mathscr{F}$, f^0, and f^- satisfy compatibility conditions of order q at the origin.*

In the same way as in the proof of the Cauchy-Kowalewsky theorem we realize that for arbitrary $\mathscr{F}$ and f^0 we can find a polynomial $\bar{u}$ such that (5.1) and (5.2) are satisfied. The equations (5.1) and (5.2), however, determine $\partial^\alpha \bar{u}(0,0)$ for $|\alpha| \leqq q$, and thus $d^\beta f^-(0)$, $\beta \leqq q$, in (5.3) cannot be chosen arbitrarily. The following theorem is easily proved.

Theorem 5.2. *There are matrices $M_{\gamma\beta}$ and $N_{\gamma\alpha}$, depending only on $\mathscr{D}$, $\mathscr{G}$, and $\mathscr{B}$, such that if $\mathscr{F}$, f^0, and f^- satisfy compatibility conditions of order q at the origin, then for $\gamma \leqq q$*

$$(5.4)\qquad d^\gamma f^-(0) = \sum_{\beta=0}^{\gamma} M_{\gamma\beta}\, d^\beta f^0(0) + \sum_{|\alpha| \leqq \gamma-1} N_{\gamma\alpha}\, \partial^\alpha \mathscr{F}(0,0),$$

and such that if on the other hand (5.4) *is satisfied for* $\mathscr{F}$, $\mathscr{f}^0$, *and* $\mathscr{f}$, *then we can find a polynomial* $\bar{u} = \bar{u}(x, t)$ *of order* q *such that* (5.1), (5.2), *and* (5.3) *are satisfied.*

In particular, for $q=1$ we have the following theorem.

Theorem 5.3. *Assume that* $\mathscr{B} \in \mathscr{C}^1_{m, m^-}(\mathscr{S})$. *Necessary conditions for the solvability in* $\mathscr{C}^1_m(\mathscr{S})$ *of the boundary problem* (1.2) *are the following relations:*

$$\mathscr{f}^-(0) = \mathscr{B}(0)\, \mathscr{f}^0(0), \tag{5.5}$$

$$d\mathscr{f}\ (0) = \mathscr{B}(0)\, [\mathscr{D}(0,0)\, d\mathscr{f}^0(0) - \mathscr{G}(0,0)\, \mathscr{f}^0(0) + \mathscr{F}(0,0)] + d\mathscr{B}(0)\, \mathscr{f}^0(0). \tag{5.6}$$

Proof. We have according to (1.2)

$$\mathscr{f}^-(0) = \mathscr{B}(0)\, u(0,0) = \mathscr{B}(0)\, \mathscr{f}^0(0),$$

$$\begin{aligned} d\mathscr{f}^-(0) &= \frac{d}{dt}\big(\mathscr{B}(t)\, u(0,t)\big)_{t=0} = \mathscr{B}(0)\, \frac{\partial u}{\partial t}(0,0) + d\mathscr{B}(0)\, u(0,0) \\ &= \mathscr{B}(0) \left[\mathscr{D}(0,0)\, \frac{\partial u}{\partial x}(0,0) - \mathscr{G}(0,0)\, u(0,0) + \mathscr{L}\, u(0,0)\right] + d\mathscr{B}(0)\, u(0,0) \\ &= \mathscr{B}(0)\, [\mathscr{D}(0,0)\, d\mathscr{f}^0(0) - \mathscr{G}(0,0)\, \mathscr{f}^0(0) + \mathscr{F}(0,0)] + d\mathscr{B}(0)\, \mathscr{f}^0(0). \end{aligned}$$

It turns out that (5.5) and (5.6) are also sufficient conditions for the solvability of the boundary problem (1.2). This will be proved later by means of Lemma 5.1. In the existence proof we consider the discrete boundary problem (2.2).

Lemma 5.1. *Assume that* $\mathscr{D}, \mathscr{G}, \mathscr{F} \in \mathscr{C}^1$; $\mathscr{B}, \mathscr{f}^0, \mathscr{f}^- \in \mathscr{C}^2$ *and that the conditions* (5.5) *and* (5.6) *are satisfied. Then there is a constant* C *such that if* u_n *is the solution of the discrete boundary problem* (2.2), *then for all* n

$$|u_n^-|_{0,2} \leqq C.$$

Proof. We have according to (2.1)

$$\Delta_\tau u^+_{\xi,\tau} = D^+_{\xi,\tau}\, \Delta^+_\xi u^+_{\xi,\tau} - G^+_{\xi,\tau}\, u_{\xi,\tau} + L^+ u_{\xi,\tau}, \qquad (\xi,\tau) = (0,0), (1,0), (0,1). \tag{5.7}$$

In particular, we obtain

$$\Delta_\tau u^+_{\xi,0} = D^+_{\xi,0}\, \Delta^+_\xi f^{0+}_\xi - G^+_{\xi,0}\, f^0_\xi + F^+_{\xi,0}, \qquad \xi = 0, 1, \tag{5.8}$$

and hence for $\xi = 0$ owing to the regularity assumptions we obtain the boundedness of $\Delta_\tau u^+_{0,0}$. Because of (2.6) for $\tau = 1, 2$ we have

$$u^-_{0,\tau} = f^-_\tau - B^+_\tau\, u^+_{0,\tau}. \tag{5.9}$$

According to (5.5), (5.9) is valid also for $\tau = 0$. Therefore we get

$$\Delta_\tau u^-_{0,0} = \Delta_\tau f^-_0 - \Delta_\tau (B^+ u^+)_{0,0}. \tag{5.10}$$

Because of the regularity assumptions, Lemma 3.2 and the already proved boundedness of $\Delta_\tau u^+_{0,0}$, the boundedness of $\Delta_\tau u^-_{0,0}$ follows from (5.10). From (5.8) we get

$$\Delta_\xi \Delta_\tau u^+_{0,0} = \Delta_\xi (D^+ \Delta_\xi f^{0+})_0 - \Delta_\xi (G^+ f^0)_0 + \Delta_\xi F^+_{0,0} \tag{5.11}$$

and hence the boundedness of $\Delta_\xi \Delta_\tau u^+_{0,0}$. From (5.7) we obtain

$$\Delta^2_\tau u^+_{0,0} = \Delta_\tau (D^+ \Delta_\xi u^+)_{0,0} - \Delta_\tau (B^+ u)_{0,0} + \Delta_\tau F^+_{0,0},$$

and hence the boundedness of $\Delta^2_\tau u^+_{0,0}$ follows from (5.11), the boundedness of $\Delta_\tau u^+_{0,0}$, and the regularity assumptions. From (5.9) we get

$$\Delta^2_\tau u^-_{0,0} = \Delta^2_\tau f^-_0 - \Delta^2_\tau (B^+ u^+)_{0,0},$$

and hence the boundedness of $\Delta^2_\tau u^-_{0,0}$ follows as above.

It remains to prove the boundedness of $\Delta_\xi \Delta_\tau u^-_{0,0}$. We have

$$\Delta_\tau u^-_{1,0} = D^-_{1,0} \Delta^-_\xi u^-_{1,0} - G^-_{1,0} u_{1,0} + L^- u_{1,0} = D^-_{1,0} \Delta_\xi f^{0-}_0 - G^-_{1,0} f^0_1 + F^-_{1,0}.$$

Together with (5.8) and (5.10), this gives after a simple computation

$$\begin{aligned}\Delta_\xi \Delta_\tau u^-_{0,0} = n\{&\mathscr{B}(0)\,[\mathscr{D}(0,0)\, d\mathscr{f}^0(0) - \mathscr{G}(0,0)\,\mathscr{f}^0(0) + \mathscr{F}(0,0)] + \\ &+ d\mathscr{B}(0)\,\mathscr{f}^0(0) - d\mathscr{f}^-(0)\} + O(1) \quad \text{when} \quad n \to \infty.\end{aligned}$$

According to (5.6) this implies the boundedness of $\Delta_\xi \Delta_\tau u^-_{0,0}$, and the proof of Lemma 5.1 is completed.

Lemma 5.2. *If* $\mathscr{F} \in \mathscr{C}^q_m(\mathscr{V})$, $\mathscr{f}^0 \in \mathscr{C}^{q+1}_m(\mathscr{S}^0)$, $\mathscr{f}^- \in \mathscr{C}^{q+1}_{m^-}(\mathscr{S}^-)$ *and if*

$$\begin{aligned} &\partial^\alpha \mathscr{F}(0,0) = 0, \qquad && |\alpha| \leqq q-1, \\ &d^\beta \mathscr{f}^0(0) \quad = 0, && \beta \leqq q, \\ &d^\beta \mathscr{f}^-(0) \quad = 0, && \beta \leqq q, \end{aligned} \tag{5.12}$$

then there is a constant C such that if u_n *is the solution of the discrete boundary problem* (2.2), *then for all* n

$$|u^-_n|_{0,q+1} \leqq C. \tag{5.13}$$

Proof. According to (2.3), (2.4), (2.5), and (2.6) we can write

$$\Delta^\alpha u^-_{0,0} = n^{|\alpha|} \sum_{\xi=1}^{|\alpha|} P_\xi f^0_\xi + n^{|\alpha|} \sum_{\tau=0}^{|\alpha|} Q_\tau f^-_\tau + n^{|\alpha|-1} \sum_{\xi+\tau \leqq |\alpha|-1} R_{\xi,\tau} F_{\xi,\tau},$$

where the coefficients P_ξ, Q_τ, and $R_{\xi,\tau}$ are bounded when $n \to \infty$. According to (5.12) we therefore have

$$\Delta^\alpha u^-_{0,0} = O(n^{|\alpha|-q-1}) = O(1) \quad \text{when} \quad n \to \infty \quad \text{for} \quad |\alpha| \leqq q+1,$$

which is (5.13).

Remark. It might be expected that in general if $\mathscr{F} \in \mathscr{C}^q_m(\mathscr{V})$, $\mathscr{f}^0 \in \mathscr{C}^{q+1}_m(\mathscr{S}^0)$, $\mathscr{f}^- \in \mathscr{C}^{q+1}_{m^-}(\mathscr{S}^-)$ satisfy compatibility conditions of order q, then (5.13) would be valid. For $q=1$, this was the case according to Lemma 5.1. For $q>1$ this is not always the case, as is proved by the following example.

Consider the boundary problem ($m=1$)

$$\begin{gathered} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \\ u(x,0) = f(x), \\ u(0,t) = g(t). \end{gathered}$$

The compatibility conditions of order 2 are here

$$g(0) = f(0),$$
$$g'(0) = -f'(0),$$
$$g''(0) = f''(0).$$

A simple computation for the solution u_n of the corresponding discrete boundary problem gives

$$\Delta_\xi^2 \Delta_\tau u_{0,0} = n^2 (f'(0) + g'(0)) + n (\lambda g''(0) - f''(0)) + O(1) \quad \text{when} \quad n \to \infty,$$
$$\Delta_\xi \Delta_\tau^2 u_{0,0} = n^2 (f'(0) + g'(0)) + n ((\lambda - 2) g''(0) + f''(0)) + O(1) \quad \text{when} \quad n \to \infty.$$

It follows that if $\lambda < 1$, then $|u_n^-|_{0,3} = O(1)$ when $n \to \infty$ only if $f''(0) = g''(0) = 0$.

6. Existence and uniqueness theorems

We first state the uniqueness theorem.

Theorem 6.1. *If* $\mathscr{F}$, $\mathscr{f}^0$, $\mathscr{f}^- \in \mathscr{C}^0$, *then there is at most one solution* $u \in \mathscr{C}_m^1(\mathscr{V})$ *of the boundary problem* (1.2).

Proof. This follows at once from Theorem 4.2.

We now prove the existence theorem in $\mathscr{C}_m^q(\mathscr{V})$ and first consider the special case $q = 1$.

Theorem 6.2. *Assume that* $\mathscr{D}$, $\mathscr{G}$, $\mathscr{B}$, $\mathscr{F}$, $\mathscr{f}^0$, $\mathscr{f}^- \in \mathscr{C}^2$ *and that the conditions* (5.5) *and* (5.6) *are satisfied. Then there is a solution* $u \in \mathscr{C}_m^1(\mathscr{V})$ *of the boundary problem* (1.2).

Proof. Consider the discrete boundary problem (2.2). Owing to Lemma 3.2 we have

$$|F^+|_{V_1,2} + |F^-|_{V_1^-,2} + |f^0|_{S^0,2} + |f^-|_{S^-,2} \leq C.$$

Lemma 5.1 gives for the solution u_n of (2.2), $|u_n^-|_{0,2} \leqq C$. Therefore according to Theorem 4.3 we have $|u_n|_{V,2} \leqq C$. Using Lemma 3.8, we can find a sequence $\{n_k\}_1^\infty$ such that $E u_{n_k}$ converges uniformly to $u \in \mathscr{C}_m(\mathscr{V})$ when $k \to \infty$ and such that $E \Delta_\xi^+ u_{n_k}^+$, $E \Delta_\xi^- u_{n_k}^-$, and $E \Delta_\tau u_{n_k}$ converge uniformly to $\partial u^+/\partial x$, $\partial u^-/\partial x$, and $\partial u/\partial t$, respectively. We assert that u is a solution of (1.2). Lemma 3.4 and Lemma 3.6 give

$$\mathscr{L}^+ u = \lim_{k\to\infty} E L^+ u_{n_k} = \lim_{k\to\infty} E R_1 \mathscr{F}^+ = \mathscr{F}^+,$$
$$\mathscr{L}^- u = \lim_{k\to\infty} E L^- u_{n_k} = \lim_{k\to\infty} E R_1^- \mathscr{F}^- = \mathscr{F}^-,$$
$$\ell^0 u = \lim_{k\to\infty} E^0 l^0 u_{n_k} = \lim_{k\to\infty} E^0 r^0 \mathscr{f}^0 = \mathscr{f}^0,$$
$$\ell^- u = \lim_{k\to\infty} E^- l^- u_{n_k} = \lim_{k\to\infty} E^- r^- \mathscr{f}^- = \mathscr{f}^-.$$

which proves Theorem 6.2.

Theorem 6.3. *Assume that* $q \geqq 1$ *and* $\mathscr{D}$, $\mathscr{G}$, $\mathscr{B}$, $\mathscr{F}$, $\mathscr{f}^0$, $\mathscr{f}^- \in \mathscr{C}^{q+1}$ *and that* $\mathscr{F}$, $\mathscr{f}^0$, $\mathscr{f}^-$ *satisfy compatibility conditions of order* q *at the origin. Then there is a solution* $u \in \mathscr{C}_m^q(\mathscr{V})$ *of the boundary problem* (1.2).

Proof. We first assume that the conditions (5.12) are satisfied. Consider the discrete boundary problem (2.2). The assumptions and Lemma 3.2 give

$$|F^+|_{V_1, q+1} + |F^-|_{V_1^-, q+1} + |f^0|_{S^0, q+1} + |f^-|_{S^-, q+1} \leqq C.$$

Lemma 5.2 gives for the solution u_n of (2.2)

$$|u_n^-|_{0, q+1} \leqq C.$$

According to Theorem 4.3 we therefore have

$$|u_n|_{V, q+1} \leqq C.$$

Using Lemma 3.8, we can find a sequence $\{n_k\}_1^\infty$ such that $E u_{n_k}$ converges uniformly to $u \in \mathscr{C}_m^q(\mathscr{V})$ when $k \to \infty$ and such that $E \Delta^\alpha u_{n_k}$ converges uniformly to $\partial^\alpha u$ when $k \to \infty$ for $|\alpha| \leqq q$. The function u is as in Theorem 6.2 a solution of (1.2).

We now assume that the conditions (5.12) are not satisfied. Since $\mathscr{F}$, $\not{f}^0$, $\not{f}^-$ fulfill compatibility conditions of order q at the origin, owing to Theorem 5.2 we can find $\bar{u} \in \mathscr{C}_m^{q+2}(\mathscr{V})$ such that (5.1), (5.2), and (5.3) are satisfied. We consider the boundary problem

$$\begin{aligned} \mathscr{L} u &= \mathscr{F} - \mathscr{L} \bar{u} \quad \text{in} \quad \mathscr{V}, \\ \ell^0 u &= \not{f}^0 - \ell^0 \bar{u} \quad \text{on} \quad \mathscr{S}^0, \\ \ell^- u &= \not{f}^- - \ell^- \bar{u} \quad \text{on} \quad \mathscr{S}^-. \end{aligned}$$

Here the conditions (5.12) are fulfilled, and thus there exists a solution $u \in \mathscr{C}_m^q(\mathscr{V})$. The function $u + \bar{u} \in \mathscr{C}_m^q(\mathscr{V})$ is then a solution of the original boundary problem (1.2). This proves Theorem 6.3.

Theorem 6.4. *Assume that* $\mathscr{D}$, $\mathscr{G}$, $\mathscr{B}$, $\mathscr{F}$, $\not{f}^0$, $\not{f}^- \in \mathscr{C}^2$, *and that the conditions* (5.5) *and* (5.6) *are satisfied. Then we have for the solution* u_n *of the discrete boundary problem* (2.2)

$$E \Delta^\alpha u_n \to \partial^\alpha u, \qquad |\alpha| \leqq 1,$$

where u *is the uniquely determined solution of the boundary problem* (1.2).

Proof. We prove the assertion for $|\alpha| = 0$. For $|\alpha| = 1$ it can be proved analogously. It follows as in the proof of Theorem 6.2 that every sequence $\{n_k\}_1^\infty$ contains a subsequence $\{n_k'\}_1^\infty$ such that $E u_{n_k'}$ converges to a solution $u_{\{n_k'\}_1^\infty}$ of the boundary problem (1.2) when $k \to \infty$. According to the uniqueness theorem (Theorem 6.1) we have $u_{\{n_k'\}_1^\infty} = u$, where u is the only solution of the boundary problem. Hence we have $E u_n \to u$ when $n \to \infty$.

It is also possible to obtain results concerning the rate of convergence, for instance the following theorem.

Theorem 6.5. *Assume that* $\mathscr{D}$, $\mathscr{G}$, $\mathscr{B}$, $\mathscr{F}$, $\not{f}^0$, $\not{f}^- \in \mathscr{C}^3$, *and that compatibility relations of order* 2 *are satisfied at the origin. If* u *and* u_n *are the solutions of the problems* (1.2) *and* (2.2), *then*

$$|E u_n - u|_{\mathscr{V}} = O(n^{-1}) \quad \textit{when} \quad n \to \infty.$$

Proof. According to Theorem 6.3, $u \in \mathscr{C}_m^2(\mathscr{V})$. We can write

$$|E u_n - u|_{\mathscr{V}} \leqq |E(u_n - R u)|_{\mathscr{V}} + |E R u - u|_{\mathscr{V}} = |u_n - R u|_V + |E R u - u|_{\mathscr{V}}.$$

According to Lemma 3.5 we have $|E R u - u|_{\mathscr{V}} = O(n^{-2})$ when $n \to \infty$. To estimate $|u_n - R u|_V$ we use Theorem 4.1 and obtain

$$\begin{aligned} |u_n - R u|_V &\leqq C\{|L^+ u_n - L^+ R u|_{V_1} + |L^- u_n - L^- R u|_{V_1^-} + \\ &\qquad + |l^0 u_n - l^0 R u|_{S^0} + |l^- u_n - l^- R u|_{S^-}\} \\ &= C\{|L^+ u_n - L^+ R u|_{V_1} + |L^- u_n - L^- R u|_{V_1^-}\}, \end{aligned}$$

since $l^0 u_n = l^0 R u = f^0$ and $l^- u_n = l^- R u = f^-$. We have $(0 \leqq \vartheta_i \leqq 1,\ i = 1, 2)$

$$\begin{aligned} |L^+ u_{n,\xi,\tau} - L^+ R u_{\xi,\tau}| &= |\mathscr{F}^+(\xi h, \tau k) - L^+ R u_{\xi,\tau}| \\ &= \left|\left(\frac{\partial u^+}{\partial t} - \mathscr{D}^+ \frac{\partial u^+}{\partial x} + \mathscr{G}^+ u - \Delta_\tau u^+ + \mathscr{D}^+ \Delta_\xi^+ u^+ - \mathscr{G}^+ u\right)_{(\xi h,\, \tau k)}\right| \\ &= \left|\left(\frac{\partial u^+}{\partial t} - \Delta_\tau u^+ - \mathscr{D}^+\left(\frac{\partial u^+}{\partial x} - \Delta_\xi^+ u\right)\right)_{(\xi h,\, \tau k)}\right| \\ &= \left|\frac{k}{2}\ \frac{\partial^2 u^+}{\partial t^2}\left(\xi h, (\tau + \vartheta_1) k\right) - \mathscr{D}^+(\xi h, \tau k)\,\frac{h}{2} \cdot \frac{\partial^2 u^+}{\partial x^2}\left((\xi + \vartheta_2) h, \tau k\right)\right| \\ &\leqq \frac{1}{2}\left(k + |\mathscr{D}^+|_{\mathscr{V}} h\right) |u^+|_{\mathscr{V},2} = O(n^{-1}) \quad \text{when} \quad n \to \infty, \end{aligned}$$

and similarly for the minus case. This proves the theorem.

7. The quasi-linear case

In this section we consider the quasi-linear boundary problem

$$\begin{aligned} &\frac{\partial u}{\partial t} - \mathscr{D}\frac{\partial u}{\partial x} = \mathscr{F}, \quad \mathscr{D} = \mathscr{D}(x, t, u), \quad \mathscr{F} = \mathscr{F}(x, t, u) \quad \text{in } \mathscr{V}, \\ (7.1) \qquad &u = \mathscr{f}^0 \quad \text{on} \quad \mathscr{S}^0, \\ &u^- = \mathscr{f}^- \quad \text{on} \quad \mathscr{S}^{0-}, \quad \mathscr{f}^- = \mathscr{f}^-(t, u^+). \end{aligned}$$

$\mathscr{V}$, $\mathscr{S}^0$, and $\mathscr{S}^-$, with λ to be specified shortly, have the same meaning as before. About $\mathscr{D}, \mathscr{F}, \mathscr{f}^0, \mathscr{f}^-$ we make the following assumptions:

1. *$\mathscr{f}^0 \in \mathscr{C}_m^2(\mathscr{S}^0)$. There are closed sets* $(u = (u_1, \ldots, u_m),\ u^+ = (u_1, \ldots, u_{m^+}))$

$$\Omega = \{(x, t, u)\,|\,0 \leqq x \leqq 1,\ 0 \leqq t \leqq \delta_1,\ |u - \mathscr{f}^0(x)| \leqq \varepsilon\},$$
$$\Omega^- = \{(t, u^+)\,|\,0 \leqq t \leqq \delta_1,\ |u^+ - \mathscr{f}^{0+}(0)| \leqq \varepsilon\}$$

such that $\mathscr{F} \in \mathscr{C}_m^2(\Omega)$, $\mathscr{D} \in \mathscr{C}_{m \times m}^2(\Omega)$, $\mathscr{f}^- \in \mathscr{C}_{m^-}^2(\Omega^-)$. *In* Ω, $\mathscr{D}$ *can be written*

$$\mathscr{D} = \begin{pmatrix} \mathscr{D}^+ & 0 \\ 0 & \mathscr{D}^- \end{pmatrix}$$

where $\mathscr{D}^+ > 0$ *and* $\mathscr{D}^- < 0$.

2. The following compatibility conditions are satisfied at the origin:

α) $\mathscr{f}^-(0, \mathscr{f}^{0+}(0)) = \mathscr{f}^{0-}(0)$,

β) $$\frac{\partial \mathscr{f}^-}{\partial t}(0, \mathscr{f}^{0+}(0)) + \frac{\partial \mathscr{f}^-}{\partial u^+}(0, \mathscr{f}^{0+}(0)) \Big[\mathscr{D}^+(0, 0, \mathscr{f}^0(0)) \frac{d\mathscr{f}^{0+}}{dx}(0) + \mathscr{F}^+(0, 0, \mathscr{f}^0(0)\Big]$$
$$= \mathscr{D}^-(0, 0, \mathscr{f}^0(0)) \frac{d\mathscr{f}^{0-}}{dx}(0) + \mathscr{F}^-(0, 0, \mathscr{f}^0(0)),$$

where $\partial\mathscr{f}^-/\partial u^+$ *is the matrix with columns* $\partial\mathscr{f}^-/\partial u_k^+$, $k = 1, \ldots, m^+$.

Under these assumptions we shall give a local existence proof for the boundary problem (7.1). Let λ be a positive number $\leqq |\mathscr{D}|_\Omega^{-1}$ and let h, k, V, *etc.* have the same meaning as before. We introduce

$$\mathscr{V}^\delta = \mathscr{V} \cap \{0 \leqq t \leqq \delta\}, \qquad V_p^\delta = V_p \cap \mathscr{V}^\delta.$$

We consider the discrete boundary problem

$$\begin{aligned} \Delta_\tau u_{\xi,\tau}^+ - D_{\xi,\tau}^+ \Delta_\xi^+ u_{\xi,\tau}^+ &= F_{\xi,\tau}^+ \quad \text{on} \quad V_1 \\ \Delta_\tau u_{\xi,\tau}^- - D_{\xi,\tau}^- \Delta_\xi^- u_{\xi,\tau}^- &= F_{\xi,\tau}^- \quad \text{on} \quad V_1^- \\ u_{\xi,0} &= f_\xi^0 \quad \text{on} \quad S^0 \\ u_{0,\tau}^- &= f_\tau^- \quad \text{on} \quad S^-. \end{aligned} \tag{7.2}$$

Here $D_{\xi,\tau}^+ = \mathscr{D}(\xi h, \tau k, u(\xi h, \tau k))$, *etc.* The following result may be obtained as in the proof of Theorem 2.1.

Theorem 7.1. *There is one and only one solution* $u \in \mathscr{C}_m(V)$ *of the discrete boundary problem* (7.2).

The solution can be defined by induction over τ by means of the following scheme:

$$\begin{aligned} & u_{\xi,0} = f_\xi^0, && \xi = 0, \ldots, n, \\ & u_{\xi,\tau+1}^+ = \lambda D_{\xi,\tau}^+ u_{\xi+1,\tau}^+ + (I^+ - \lambda D_{\xi,\tau}^+) u_{\xi,\tau}^+ + k F_{\xi,\tau}^+, && \xi = 0, \ldots, n - \tau - 1, \\ & u_{\xi,\tau+1}^- = (I^- + \lambda D_{\xi,\tau}^-) u_{\xi,\tau}^- - \lambda D_{\xi,\tau}^- u_{\xi-1,\tau}^- + k F_{\xi,\tau}^-, && \xi = 1, \ldots, n - \tau - 1, \\ & u_{0,\tau+1}^- = f_{\tau+1}^-. \end{aligned} \tag{7.3}$$

Theorem 7.2. *Assume that* $\mathscr{D}, \mathscr{F}, \mathscr{f}^0, \mathscr{f}^-$ *satisfy 1. and 2. Then if u is the solution of the discrete boundary problem* (7.2), *there is a* $\delta > 0$ *and a constant* C *such that independently of* n, $|u|_{V^\delta, 2} \leqq C$.

For the proof we need Lemma 7.2, which is proved by means of Lemma 7.1.

Lemma 7.1. *Assume that* $a_\tau \geqq 0$, $\tau = 0, 1, \ldots, n$, *and*

$$a_{\tau+1} \leqq a_\tau + n^{-1} f(a_\tau), \qquad \tau = 0, \ldots, n - 1,$$

where $f(x)$ *is non-decreasing, continuous, and* > 0 *for* $x \geqq 0$. *Then there exist a* $\delta > 0$ *and a continuous, non-decreasing function* $\psi(x)$ *defined for* $x \geqq 0$ *such that* $a_\tau \leqq \psi(a_0)$ *for* $\tau n^{-1} \leqq \delta$.

Proof. We put $b_\tau = a_0 + 2\tau n^{-1} f(a_0)$. We have $b_0 = a_0$ and

$$b_{\tau+1} - b_\tau - n^{-1} f(b_\tau) = n^{-1}[2f(a_0) - f(a_0 + 2f(a_0)\,\tau n^{-1})]. \tag{7.4}$$

Consider the continuous function $\chi(z) = 2f(a_0) - f(a_0 + 2f(a_0)\,z)$. We have $\chi(0) = f(a_0) > 0$, and therefore there exists a $\delta > 0$ such that $\chi(z) \geqq 0$ for all z with $0 \leqq z \leqq \delta$. With this δ we have according to (7.4)

$$b_{\tau+1} \geqq b_\tau + n^{-1} f(b_\tau), \qquad \tau n^{-1} \leqq \delta. \tag{7.5}$$

Since $f(x)$ is non-decreasing, (7.5) implies $a_\tau \leqq b_\tau$, and therefore with $\psi(x) = x + 2\delta f(x)$,

$$a_\tau \leqq a_0 + 2\delta f(a_0) = \psi(a_0), \qquad \tau n^{-1} \leqq \delta.$$

Lemma 7.2. *Assume that* $a_\tau^+ \geqq 0$, $a_\tau^- \geqq 0$, $a_\tau = \max(a_\tau^+, a_\tau^-)$, $\tau = 0, \ldots, n$, *and that*

$$\begin{aligned} a_{\tau+1}^+ &\leqq a_\tau^+ + n^{-1} f(a_\tau), && \tau = 0, 1, \ldots, n-1, \\ a_{\tau+1}^- &\leqq \max\{a_\tau^- + n^{-1} f(a_\tau),\, g(a_{\tau+1}^+)\}; && \tau = 0, 1, \ldots, n-1, \end{aligned}$$

where $f(x)$ *is non-decreasing, continuous and* >0 *for* $x \geqq 0$ *and* $g(x)$ *is non-decreasing and continuously differentiable with* $g'(x) \geqq 1$ *for* $x \geqq 0$. *Then there exist a* $\delta > 0$ *and a continuous, non-decreasing function* $\psi(x)$ *defined for* $x \geqq 0$ *such that* $a_\tau \leqq \psi(a_0)$ *for* $\tau n^{-1} \leqq \delta$.

Proof. We put

$$\begin{aligned} b_0^\pm &= a_0^\pm \\ b_\tau &= \max(b_\tau^+, b_\tau^-), && \tau = 0, 1, \ldots, n, \\ b_{\tau+1}^+ &= b_\tau^+ + n^{-1} f(b_\tau), && \tau = 0, 1, \ldots, n-1, \\ b_{\tau+1}^- &= \max\{b_\tau^- + n^{-1} f(b_\tau),\, g(b_{\tau+1}^+)\}, && \tau = 0, 1, \ldots, n-1. \end{aligned} \tag{7.6}$$

Since f and g are non-decreasing it follows easily that $a_\tau^\pm \leqq b_\tau^\pm$, $a_\tau \leqq b_\tau$. It is therefore sufficient to prove that there exist a $\delta > 0$ and a continuous, non-decreasing function $\psi(x)$ such that $b_\tau \leqq \psi(b_0)$ for $\tau n^{-1} \leqq \delta$. Suppose that for some τ in (7.6), $b_\tau^- = g(b_\tau^+)$. Then we also have $b_{\tau+1}^- = g(b_{\tau+1}^+)$ for $(0 \leqq \vartheta \leqq 1)$

$$g(b_{\tau+1}^+) = g(b_\tau^+ + n^{-1} f(b_\tau)) = g(b_\tau^+) + n^{-1} f(b_\tau)\, g'(b_\tau^+ + \vartheta n^{-1} f(b_\tau)) \geqq b_\tau^- + n^{-1} f(b_\tau).$$

It follows that there is a τ_0 such that

$$\begin{aligned} b_{\tau+1}^- &= b_\tau^- + n^{-1} f(b_\tau), && \tau < \tau_0, \\ b_{\tau+1}^- &= g(b_{\tau+1}^+), && \tau \geqq \tau_0. \end{aligned} \tag{7.7}$$

The case where $b_\tau^- \neq g(b_\tau^+)$ for $\tau = 0, \ldots, n$ also fits in here if we agree to put $\tau_0 = n$ in this case. For $\tau < \tau_0$ we thus have

$$b_{\tau+1} \leqq b_\tau + n^{-1} f(b_\tau).$$

According to Lemma 7.1 there exist a $\delta_1 > 0$ and a continuous, non-decreasing function $\psi_1(x)$ such that

$$b_\tau \leqq \psi_1(b_0), \qquad \tau n^{-1} \leqq \delta_1, \qquad \tau \leqq \tau_0. \tag{7.8}$$

We have

$$b^+_{\tau_0+1} = b^+_{\tau_0} + n^{-1} f(b_{\tau_0}) \leqq b_{\tau_0} + f(b_{\tau_0}) \leqq \psi_1(b_0) + f\big(\psi_1(b_0)\big) = \psi_2(b_0). \tag{7.9}$$

For $\tau > \tau_0$ we have

$$b^+_{\tau+1} = b^+_\tau + n^{-1} f(b_\tau) \leqq b^+_\tau + n^{-1} f\big(\max(b^+_\tau, g(b^+_\tau))\big).$$

The function $f\big(\max(x, g(x))\big)$ is non-decreasing, and therefore according to Lemma 7.1 and (7.9) there exist a $\delta_2 > 0$ and continuous, non-decreasing functions $\psi_3(x)$ and $\psi_4(x)$ such that

$$b^+_\tau \leqq \psi_3(b^+_{\tau_0+1}) \leqq \psi_3\big(\psi_2(b_0)\big) = \psi_4(b_0); \qquad \tau n^{-1} \leqq \delta_2, \quad \tau > \tau_0. \tag{7.10}$$

Finally, (7.7), (7.8), and (7.10) imply with $\delta = \min(\delta_1, \delta_2)$ and $\psi(x) = \max\big(\psi_1(x), \psi_4(x), g(\psi_4(x))\big)$ that

$$b_\tau \leqq \psi(b_0), \qquad \tau n^{-1} \leqq \delta,$$

which completes the proof of Lemma 7.2.

Proof of Theorem 7.2. From the difference scheme (7.3) one easily deduces the following representation of the successive differences with respect to τ:

$$\begin{aligned} \Delta^j_\tau u^+_{\xi,\tau+1} &= \lambda D^+_{\xi,\tau+j} \Delta^j_\tau u^+_{\xi+1,\tau} + (I^+ - \lambda D^+_{\xi,\tau+j}) \Delta^j_\tau u^+_{\xi,\tau} + k\, Q^+_{j,\xi,\tau}, \\ &\qquad\qquad \xi = 0, 1, \ldots, n - \tau - j - 1; \\ \Delta^j_\tau u^-_{\xi,\tau+1} &= (I^- + \lambda D^-_{\xi,\tau+j}) \Delta^j_\tau u^-_{\xi,\tau} - \lambda D^-_{\xi,\tau+j} \Delta^j_\tau u^-_{\xi-1,\tau} + k\, Q^-_{j,\xi,\tau}, \\ &\qquad\qquad \xi = 1, \ldots, n - \tau - j - 1; \\ \Delta^j_\tau u^-_{0,\tau+1} &= \Delta^j_\tau f^-_{\tau+1}. \end{aligned} \tag{7.11}$$

Here $Q^\pm_{0,\xi,\tau} = F^\pm_{\xi,\tau}$, and

$$Q^\pm_{j+1,\xi,\tau} = \Delta_\tau Q^\pm_{j,\xi,\tau} + \Delta_\tau D^\pm_{\xi,\tau+j} \Delta^j_\tau \big\{(D^\pm_{\xi,\tau})^{-1} (\Delta_\tau u^\pm_{\xi,\tau} - F^\pm_{\xi,\tau})\big\}.$$

In the scheme (7.3) we consider only ξ, τ such that $(\xi h, \tau k, u_{\xi,\tau}) \in \Omega$, that is, ξ, τ such that $0 \leqq \tau k \leqq \delta_1$; $0 \leqq \xi h \leqq 1 - \tau k$; $|u_{\xi,\tau} - f^0_\xi| \leqq \varepsilon$. It is then easy to prove that

$$\begin{aligned} |Q^\pm_{1,\xi,\tau}| &\leqq C_1\big(|\Delta_\tau u|^2_{S^\tau_1} + 1\big), \\ |Q^\pm_{j,\xi,\tau}| &\leqq C_j\big(|\Delta^j_\tau u|_{S^\tau_j} + 1\big), \qquad j > 1, \end{aligned} \tag{7.12}$$

where C_j is bounded with $|\Delta^i_\tau u|_{S^\tau_i}$, $i = 0, \ldots, j-1$. (In fact we shall not use these estimates for $j > 2$.) Therefore from (7.11) and (7.12) we find that

$$\begin{aligned} |\Delta_\tau u^+|_{S^{\tau+1}_1} &\leqq |\Delta_\tau u^+|_{S^\tau_1} + k\, C_1\big(|\Delta_\tau u|^2_{S^\tau_1} + 1\big), \\ |\Delta_\tau u^-|_{S^{\tau+1}_1} &\leqq \max\big\{|\Delta_\tau u^-|_{S^\tau_1} + k\, C_1\big(|\Delta_\tau u|^2_{S^\tau_1} + 1\big),\ K_1\big(|\Delta_\tau u^+|_{S^{\tau+1}_1} + 1\big)\big\}, \end{aligned}$$

where the assumption that the constant K_1 satisfies $K_1 \geqq 1$ is no restriction. We shall see now that $|\Delta_\tau u|_{S_1^0}$, i.e., $\Delta_\tau u^{\pm}_{\xi,0}$ is bounded. For the $+$ case and $\xi = 0, 1, \ldots$ and for the $-$ case and $\xi = 1, 2, \ldots$ this follows at once from (7.2) and assumption *1*. Further we have in view of the compatibility condition *2*α

$$\Delta_\tau u^-_{0,0} = \Delta_\tau f^-_0,$$

which is bounded with $\Delta_\tau u^+_{0,0}$. Lemma 7.2 now shows that there exist a constant C and a $\delta_2 > 0$, $\delta_2 \leqq \delta_1$, such that

$$|\Delta_\tau u|_{S_1^\tau} \leqq C, \qquad \tau k \leqq \delta_2.$$

The assumption $|u_{\xi,\tau} - f^0_\xi| \leqq \varepsilon$ for $\tau \leqq \tau_0$, where $\tau_0 k \leqq \delta_2$, thus implies $|\Delta_\tau u|_{S_1^\tau} \leqq C$ for $\tau \leqq \tau_0$. It is therefore an easy induction to prove that

$$|u_{\xi,\tau} - f^0_\xi| \leqq C \cdot \delta_3 \leqq \varepsilon, \qquad \tau k \leqq \delta_3 = \min\left(\frac{\varepsilon}{C}, \delta_2\right).$$

From (7.2) and (4.13) we can estimate $\Delta_\xi u_{\xi,\tau}$, and we thus find that there exist a $\delta > 0$ and a C such that $|u|_{V^{\delta},1} \leqq C$. It remains to estimate the second-order differences. We consider $\Delta_\tau^2 u_{\xi,\tau}$ and first observe that $\Delta_\tau^2 u_{\xi,0}$ is bounded. For $\Delta_\tau^2 u^+_{\xi,0}$, $\xi = 0, 1, \ldots$ and $\Delta_\tau^2 u^-_{\xi,0}$, $\xi = 2, 3, \ldots$ this follows from (7.2). Also $\Delta_\tau^2 u^-_{0,0} = \Delta_\tau^2 f^-_0$ is bounded with $\Delta_\tau^j u^+_{0,0}$, $j = 0, 1, 2$. Finally using (7.2), we find that $\Delta_\tau^2 u^-_{1,0}$ is bounded with $\Delta_\xi \Delta_\tau u^-_{0,0}$. We have easily from the definitions

$$\Delta_\xi \Delta_\tau u^-_{0,0} = h^{-1}\Big\{\mathscr{D}^-\big(0,0,\mathscr{f}^0(0,0)\big)\frac{d\mathscr{f}^{0-}}{dx}(0) + \mathscr{F}^-\big(0,0,\mathscr{f}^0(0)\big) - \frac{\partial \mathscr{f}^-}{\partial t}\big(0,\mathscr{f}^{0+}(0)\big) -$$
$$- \frac{\partial \mathscr{f}^-}{\partial u^+}\big(0,\mathscr{f}^{0+}(0)\big)\Big[\mathscr{D}^+\big(0,0,\mathscr{f}^0(0)\big)\frac{d\mathscr{f}^{0+}}{dx}(0) + \mathscr{F}^+\big(0,0,\mathscr{f}^0(0)\big)\Big]\Big\} + O(1) \text{ when } n \to \infty,$$

and thus according to the compatibility condition *2*β, $\Delta_\xi \Delta_\tau u^-_{0,0} = O(1)$ when $n \to \infty$. We have with a constant K_2

$$|\Delta_\tau^2 u^+|_{S_2^{\tau+1}} \leqq |\Delta_\tau^2 u^+|_{S_2^\tau} + k C_2\big(|\Delta_\tau^2 u|_{S_2^\tau} + 1\big),$$
$$|\Delta_\tau^2 u^-|_{S_2^{\tau+1}} \leqq \max\big\{|\Delta_\tau^2 u^-|_{S_2^\tau} + k C_2\big(|\Delta_\tau^2 u|_{S_2^\tau} + 1\big),\ K_2\big(|\Delta_\tau^2 u^+|_{S_2^{\tau+1}} + 1\big)\big\},$$

and therefore, since $|\Delta_\tau^2 u|_{S_2^0}$ is bounded according to Lemma 4.2, $|\Delta_\tau^2 u|_{V_2^\delta}$ is bounded. In the same way as in Section 4 we establish the boundedness of $|\Delta_\xi^2 u|_{V_2^\delta}$ and $|\Delta_\xi \Delta_\tau u|_{V_2^\delta}$, and thus Theorem 7.2 is proved.

As a consequence of Theorem 7.2 we have the following theorem, which is proved in the same way as Theorem 6.2.

Theorem 7.3. *Assume that* $\mathscr{D}, \mathscr{F}, \mathscr{f}^0, \mathscr{f}$ *satisfy 1. and 2. Then there exists a* $\delta > 0$ *such that the boundary problem* (7.1) *has a solution in* $\mathscr{V}^\delta$.

We supplement the existence theorem by means of the following uniqueness theorem.

Theorem 7.4. *Assume that* $\mathscr{f}^0 \in \mathscr{C}^0_m(\mathscr{S}^0)$, *that*

$$\Omega = \{(x,t,u) \mid 0 \leqq x \leqq 1,\ 0 \leqq t \leqq \delta_1,\ |u - \mathscr{f}^0(x)| \leqq \varepsilon\},$$
$$\Omega^- = \{(t,u^+) \mid 0 \leqq t \leqq \delta_1,\ |u^+ - \mathscr{f}^{0+}(0)| \leqq \varepsilon\},$$

and that $\mathscr{F} \in \mathscr{C}^1_m(\Omega)$, $\mathscr{D} \in \mathscr{C}^1_{m\times m}(\Omega)$, $\not{f}^- \in \mathscr{C}^1_{m^-}(\Omega^-)$, *and*

$$\mathscr{D} = \begin{pmatrix} \mathscr{D}^+ & 0 \\ 0 & \mathscr{D}^- \end{pmatrix}$$

in Ω, *where* $\mathscr{D}^+ > 0$ *and* $\mathscr{D}^- < 0$. *Assume further that* u *and* v *are solutions in* $\mathscr{C}^1_m(\mathscr{V}^{\delta_2})$ *of the boundary problem* (7.1). *Then there exists a* $\delta > 0$ *such that* $u = v$ *in* $\mathscr{V}^\delta$.

Proof. Let $\delta_3 > 0$ be so small that $(x, t, u(x,t)) \in \Omega$ and $(x, t, v(x,t)) \in \Omega$ for all $(x,t) \in \mathscr{V}^{\delta_3}$. Let $\mathscr{L} w = \frac{\partial w}{\partial t} - \mathscr{D}(u)\frac{\partial w}{\partial x}$ and $\ell^- w = w^- \big(w = (w^+, w^-)\big)$. Then $w = u - v$ satisfies the linear boundary problem $\big(\mathscr{F}(u) = \mathscr{F}(x, t, u(x,t)), \not{f}^-(u^+) = \not{f}^-(t, u^+(0,t))\big)$

$$\begin{aligned} &\mathscr{L} w = \mathscr{F}(u) - \mathscr{F}(v) + [\mathscr{D}(u) - \mathscr{D}(v)]\frac{\partial v}{\partial x}, \\ &\ell^0 w = 0, \\ &\ell^- w = \not{f}^-(u^+) - \not{f}^-(v^+). \end{aligned}$$

According to Lemma 4.4 we have for $\delta < \delta_3$

$$|w^+|_{\mathscr{V}^\delta} \leqq C\,\delta \left\{ \left| \mathscr{F}(u) - \mathscr{F}(v) + [\mathscr{D}(u) - \mathscr{D}(v)]\frac{\partial v}{\partial x}\right|_{\mathscr{V}^\delta} + |\not{f}^-(u^+) - \not{f}^-(v^+)|_{(\mathscr{S}^-)^\delta} \right\},$$

and thus with

$$C_1 = m\left|\frac{\partial \mathscr{F}}{\partial u}\right|_\Omega + m\left|\frac{\partial \mathscr{D}}{\partial u}\right|_\Omega \left|\frac{\partial v}{\partial x}\right|_{\mathscr{V}^{\delta_3}} + m^+\left|\frac{\partial \not{f}^-}{\partial u^+}\right|_{\Omega^-},$$

where the definitions of the matrices $\partial\mathscr{F}/\partial u$ and $\partial\mathscr{D}/\partial u$ are obvious, we get

$$|w^+|_{\mathscr{V}^\delta} \leqq C\,C_1\,\delta\,|w|_{\mathscr{V}^\delta}, \qquad \delta < \delta_3. \tag{7.13}$$

Further, in the same way for $\delta < \delta_3$ Lemma 4.4 yields

$$|w^-|_{\mathscr{V}^\delta} \leqq C\,C_1\,\delta\,|w|_{\mathscr{V}^\delta} + C\,C_1\,|w^+|_{\mathscr{V}^\delta},$$

and thus using (7.13), we have

$$|w^-|_{\mathscr{V}^\delta} \leqq \big(C\,C_1 + (C\,C_1)^2\big)\,\delta\,|w|_{\mathscr{V}^\delta}, \qquad \delta < \delta_3,$$

and with $C_2 = C\,C_1 + (C\,C_1)^2$

$$|w|_{\mathscr{V}^\delta} \leqq C_2\,\delta\,|w|_{\mathscr{V}^\delta}.$$

Now if $\delta < \min(\delta_3, C_2^{-1})$, we obtain $|w|_{\mathscr{V}^\delta} = 0$, which proves the theorem.

References

[1] Campbell, L. L., & A. Robinson: Mixed problems for hyperbolic partial differential equations. Proc. London Math. Soc. (3) **5**, 129—147 (1955).

[2] Courant, R., E. Isaacson & M. Rees: On the solution of nonlinear hyperbolic differential equations by finite differences. Comm. Pure Appl. Math. **5**, 243—255 (1952).

[3] Courant, R., & P. D. Lax: Nonlinear partial differential equations with two independent variables. Comm. Pure Appl. Math. **2**, 255—273 (1949).

[4] Duff, G. F. D.: Mixed problems for linear systems of first order equations. Canad. J. Math. **10**, 127—160 (1958).

[*5*] KREISS, H.-O.: Über die Lösung von Anfangsrandwertaufgaben für partielle Differentialgleichungen mit Hilfe von Differenzengleichungen. Kungl. Tekn. Högsk. Handl. Stockholm, No. 166, 61 pp. (1960).

[*6*] LAX, P. D., & J. B. KELLER: The initial and mixed boundary value problem for hyperbolic systems. Los Alamos Scientific Laboratory unclassified report LAMS-1205 (1951).

[*7*] PETROWSKI, I. G.: Vorlesungen über partielle Differentialgleichungen. Leipzig 1955.

[*8*] PEYSER, G.: Energy integrals for the mixed problem in hyperbolic partial differential equations of higher order. J. Math. Mech. **6**, 641–653 (1957).

[*9*] RICHTMYER, R. D.: Difference Methods for Initial-Value Problems. New York 1957.

[*10*] THOMÉE, V.: Estimates of the Friedrichs-Lewy type for mixed problems in the theory of linear hyperbolic differential equations in two independent variables. Math. Scand. **5**, 93–113 (1957).

[*11*] THOMÉE, V.: Existence proofs for mixed problems for hyperbolic differential equations in two independent variables by means of the continuity method Math. Scand. **6**, 5–32 (1958).

Department of Mathematics
University of Stockholm

(Received February 28, 1961)

EDITORIAL BOARD

H. BECKERT
Mathematisches Institut
der Universität Leipzig

R. BERKER
Mühürdar Caddesi No. 104
Kadiköy, Istanbul

L. CESARI
University of Michigan
Ann Arbor, Michigan

L. COLLATZ
Institut für Angewandte Mathematik
Universität Hamburg

A. ERDÉLYI
California Institute of Technology
Pasadena, California

J. L. ERICKSEN
The Johns Hopkins University
Baltimore, Maryland

G. FICHERA
Istituto Matematico
Università di Roma

R. FINN
Stanford University
California

HILDA GEIRINGER
Harvard University
Cambridge, Massachusetts

H. GÖRTLER
Institut für Angewandte Mathematik
Universität Freiburg i. Br.

D. GRAFFI
Istituto Matematico „Salvatore Pincherle"
Università di Bologna

A. E. GREEN
King's College
Newcastle-upon-Tyne

J. HADAMARD
Institut de France
Paris

L. HÖRMANDER
Department of Mathematics
University of Stockholm

M. KAC
Rockefeller Institute
New York 21, New York

T. KATO
University of Tokyo
Bunkyo-ku, Tokyo

E. LEIMANIS
University of British Columbia
Vancouver

C. C. LIN
Massachusetts Institute of Technology
Cambridge, Massachusetts

W. MAGNUS
Institute of Mathematical Sciences
New York University
New York City

G. C. McVITTIE
University of Illinois Observatory
Urbana, Illinois

J. MEIXNER
Institut für Theoretische Physik
Technische Hochschule Aachen

C. MÜLLER
Institut für Reine u. Angewandte
Mathematik
Technische Hochschule Aachen

W. NOLL
Carnegie Institute of Technology
Pittsburgh, Pennsylvania

A. OSTROWSKI
Certenago-Montagnola
Ticino

R. S. RIVLIN
Division of Applied Mathematics
Brown University
Providence, Rhode Island

M. M. SCHIFFER
Stanford University
California

J. SERRIN
Institute of Technology
University of Minnesota
Minneapolis, Minnesota

E. STERNBERG
Division of Applied Mathematics
Brown University
Providence, Rhode Island

R. A. TOUPIN
Naval Research Laboratory
Washington 25, D.C.

C. TRUESDELL
The Johns Hopkins University
Baltimore, Maryland

H. VILLAT
47, bd. A. Blanqui
Paris XIII

FILED IN STACKS

CONTENTS

ERICKSEN, J. L., Poiseuille Flow of Certain Anisotropic Fluids 1

ADKINS, J. E., & A. E. GREEN, The Finite Flexure of an Aeolotropic Elastic Cuboid 9

HAYES, M., & R. S. RIVLIN, Propagation of a Plane Wave in an Isotropic Elastic Material Subjected to Pure Homogeneous Deformation 15

MASAITIS, Č., On the Motion of Two Linked Bodies 23

LEE, E. B., & L. MARKUS, Optimal Control for Nonlinear Processes . . 36

BROWNELL, F. H., Finite Dimensionality of the Schrödinger Operator Bottom 59

THOMÉE, V., Difference Methods for Two-Dimensional Mixed Problems for Hyperbolic First Order Systems 68

Druck der Universitätsdruckerei H. Stürtz AG., Würzburg
Printed in Germany